МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное
учреждение высшего образования
Санкт-Петербургский государственный технологический институт
(технический университет)

Кафедра системного анализа и информационных технологий

А.А. Мусаев

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Лекции

Учебное пособие

Санкт-Петербург
2018

УДК 681.3.06

Мусаев, А.А. Интеллектуальный анализ данных: учебное пособие. /
А.А.Мусаев – СПб.: СПбГТИ(ТУ), 2018. – 56 с.

В учебном пособии рассматривается область системного анализа, связанная с получением знаний из больших массивов структуризированных данных. Пособие состоит из трех разделов. Первый раздел посвящен общим вопросам моделирования и системного анализа, связанным с извлечением знаний из информации, хранящейся в базах данных. Во втором разделе рассмотрены статистические методы анализа данныхю В третьем разделе приведены алгоритмы компьютерной математики, связанные с саморазвивающимися алгоритмами анализа данных – нейросетевые, эволюционные, генетические и т.п.

Учебное пособие предназначено для студентов очной формы обучения по направлению подготовки 27.03.03 «Системный анализ и управление» в рамках рабочей программы дисциплины «Системный анализ, оптимизация и принятие решений» и «Управление в организационных системах».

Учебное пособие формирует компетенции: **ОПК-1** в части – способен применять методы математики, теории управления и системного анализа, **ОПК-2** в части – способен применять аналитические, вычислительные и системно-аналитические методы для решения прикладных задач системного анализа, **ПК-4** в части – способен применять методы системного анализа для решения прикладных проектно-конструкторских задач.

Данное учебное пособие полезно студентам заочного отделения.

Рис. 2, табл. 45, формул 21, библиогр. 6 назв.

Рецензенты:

1 Государственная полярная академия, доцент кафедры математического моделирования социально-экономических и природных процессов, канд. физ-мат. наук, Валентин Гавриилович Никитенко.

2 Полосин Андрей Николаевич, канд. техн. наук, доцент кафедры систем автоматизированного проектирования и управления СПбГТИ(ТУ).

Издание подготовлено в рамках выполнения государственного задания по оказанию образовательных услуг Минобрнауки России.

Утверждено на заседании учебно-методической комиссии факультета Информатики и Управления 06.12.2017.

Рекомендовано к изданию РИС СПбГТИ(ТУ)

I. СИСТЕМОЛОГИЯ АНАЛИЗА ДАННЫХ

ЛЕКЦИЯ 1. ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ

1. Информация, данные, знания

Основными составляющими окружающего нас мира являются материя, энергия и информация. Сразу же возникает очень непростой вопрос: Что такое информация?

В соответствии с ГОСТ 7.0 – 99 имеем:

Информация – это сведения, воспринимаемые человеком или специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации.	**NB!**

Замечание. Здесь и далее аббревиатура «Nota bene», т.е. «Заметь хорошо!» будут означать пожелание автора особенно внимательно отнестись к отмеченному тексту.

Можно привести и другие определения, используемые философами, математиками, специалистами в области информатики и др.:

- Информация есть отражение реального мира: это сведения, которые один реальный объект содержит о другом реальном объекте (материалистическая философия);

- Информация – это сведения независимо от формы их представления (https://ru.wikipedia.org/wiki/Информация);

- Информация – это знания относительно фактов, событий, вещей, идей и понятий, которые в определённом контексте имеют конкретный (ISO/IEC 2382:2015).

Насколько удачны приведенные определения информации?

Очевидно, что понятнее от этих определений смысл термина не становиться – термин информация заменяется своими синонимами – «сведения», «знания» и т.п., которые, собственно, и являются информацией. Гностический змей очередной раз схватил сам себя за хвост.

Сама по себе информация может быть отнесена к категории абстрактных понятий типа математических, но ряд особенностей приближает ее к материальным объектам. Так, информацию можно получить, записать, удалить, передать; информация не может возникнуть из ничего.

Однако при распространении информации проявляется такое ее свойство, которое не присуще материальным объектам – отсутствие закона сохранения.

При передаче информации из одной системы в другую количество информации в передающей системе не уменьшится, хотя в принимающей системе оно обычно увеличивается.

Если бы информация не обладала этим свойством, то, например, преподаватель, читая лекцию студентам, терял бы информацию и становился неучем.

Бернард Шоу сказал: «Если мы с вами обменяемся яблоками, у каждого будет по одному яблоку; если обменяемся идеями – у каждого будет по две идеи».

Интересный подход к косвенному определению информации может быть сформирован из анализа динамики диалектической пары понятий «Порядок – Хаос».

В этом контексте информация представляет собой степень упорядоченности первоначально хаотической среды, т.е. величину, обратную к энтропии. Иными словами, именно информация выводит мир из состояния теплового хаоса, внедряя в него разнородные закономерности и организованные системы.

Переходя от проблем, связанных со сложностью определений первопричинных категорий к инженерной конкретике предметной области, введем общепринятые определения *данных* и *знаний*.

Данные могут представлять собой факты, понятия или команды, представленные в формализованном виде, позволяющем осуществить их передачу, интерпретацию или обработку [Обработка данных. Словарь. Основные термины. – 1992].

Данными могут быть цифровые массивы, факты, тексты, графики, картинки, звуки, аналоговые или цифровые видеосегменты. Они может быть получены в результате наблюдений, измерений, экспериментов, арифметических и логических операций.

Много ли в мире накоплено цифровых данных? В 2015-16 г.г. объем цифровой информации, созданной человечеством составил 4 зетабайт данных. К 2020г. ожидается увеличение этого объема до 40 зетабайт.

Справка:

Мегабайты: 1 Мбт = 10^6 бт;
Гигабайты: 1 Гбт = 10^9 бт;
Терабайты: 1 Тбт = 10^{12} бт;
Петабайты: 1 Пбт = 10^{15} бт;
Эксобайты: 1 Эбт = 10^{18} бт;
Зеттабайты: 1 Збт = 10^{21} бт.

При этом 90% накопленной информации было создано в течение последних 5 лет. Это и называется *информационным взрывом!*

За 2002 год, согласно оценке, сделанной в калифорнийском университете Berkeley, объем информации в мире увеличился на $5*10^{18}$ байт и удваивается каждые 2 года.

Примечание. Если предположить, что один байт соответствует песчинке, количество песчинок в зеттабайте соответствует по объему числу песчинок в плотинах 500 самых крупных в мире гидроэлектростанций.

Одна электронная книга объемом в 1000 страниц – это примерно 3 мегабайта, так что библиотека, содержащая все эти 2.73 тб информации, должна содержать примерно один квадриллион томов.

В то же время, известно, что сейчас в мире имеется примерно 130 миллионов книг. Оценить их информационную емкость чрезвычайно трудно. Предположительно имеется около 10 миллионов томов, действительно содержащих полезные сведения. При записи в электронной форме они займут 30 Птб, т.е. в миллион раз меньше всей накопленной информации. Это иллюстрирует разницу между первичной информацией и результатом ее перевода в концентрированную форму – знаниями.

Примечание. Сверхбольшая база данных (*Very Large Database, VLDB*) — это база данных, которая занимает чрезвычайно большой объём на устройстве физического хранения. Термин подразумевает максимально возможные объёмы БД, которые определяются последними достижениями в технологиях физического хранения данных и в технологиях программного оперирования данными.

Количественное определение понятия «чрезвычайно большой объём» меняется во времени. Так, в 1997 году самой большой в мире была текстовая база данных Knight Ridder's DIALOG объёмом 7 терабайт. В 2001 году самой большой считалась база данных объёмом 10,5 терабайт, в 2003 году — объёмом 25 терабайт. В 2005 году самыми крупными в мире считались базы данных с объёмом хранилища порядка сотни терабайт. В 2006 году поисковая машина Google использовала базу данных объёмом 850 терабайт

К 2010 году считалось, что объём сверхбольшой базы данных должен измеряться по меньшей мере петабайтами.

К 2014 году по косвенным оценкам компания Google хранила на своих серверах до 10—15 экзабайт данных в совокупности.

По некоторым оценкам, к 2025 году генетики будут располагать данными о геномах от 100 миллионов до 2 миллиардов человек, и для хранения подобного объёма данных потребуется от 2 до 40 экзабайт.

Итак, в руках аналитиков оказываются огромные массивы данных, полученных в результате мониторинга производственных процессов, накопленных в процессе массовых медицинских диспансеризаций или собранных в процессе социологических опросов. И что толку? Эти данные сами по себе бесполезны, их необходимо обработать, чтобы извлечь из них полезные знания.

Знания – это факты и правила, формализующие опыт специалистов в конкретной предметной области и позволяющие давать ответы (решения), которые не содержатся в исходной информации в явном виде.

NB!

Отсюда возникает *главная задача анализа данных* – преобразование данных в знание, т.е. в особый вид доступной для человеческого понимания информации.

Сами по себе знания не нужны, однако они могут быть реализованы для формирования управляющих решений.

NB! Соответственно, *целью анализа данных* является повышение качества (или обоснованности) формируемых управляющих решений.

В свою очередь качество решений оценивается через эффективность систем управления и их реализаций в конкретных предметных областях.

Например, в промышленности терминальное качество управляющих решений оценивается, как правило, в терминах экономической эффективности предприятия.

2. Три уровня анализа информации

В свете идей *семиотики* (науки о знаковых системах) адекватность информации, соответствие ее содержания образу отображаемого объекта, может выражаться в трех формах:

- синтаксической;
- семантической;
- прагматической.

Синтаксическая адекватность связана с воспроизведением формально-структурных характеристик отражения, абстрагирование от смысловых и потребительских параметров. На синтаксическом уровне учитываются: *тип носителя, способ представления, скорость передачи и обработки, формат кодов, надежность и точность преобразования* и т.п. При этом информация *инвариантна* по отношению к энергетическим и пространственно-временным свойствам своего носителя. Одна и та же информация может существовать в различных кодах.

Семантическая форма обеспечивает формирование понятий и представлений, выявление смысла, содержания информации. Количество семантической информации в сообщении является величиной относительной: одно и то же сообщение может иметь смысловое содержание для компетентного пользователя и быть бессмысленным (семантическим шумом) для пользователя некомпетентного.

Прагматический аспект рассмотрения информации связан с ее ценностью, полезностью, практическим использованием для достижения целей деятельности системы.

Возможность и эффективность использования информации обусловливается такими ее потребительскими показателями качества, как *репрезентативность, содержательность, достаточность, доступность, своевременность, устойчивость, точность, достоверность, актуальность и ценность.*

Кроме того, информация характеризуется такими свойствами, как *относительность, структурированность, наличие связи с носителем, инвариантность, содержательность, преобразуемость, совместимость, надежность, избыточность, защищенность* и другие.

Всю информацию, поступающую в систему анализа и находящуюся в ней, можно подразделить на

- ***процедурную**, согласно которой реализуется процесс обработки, и*
- ***декларативную*** информацию, подвергающуюся обработке. При этом процедурная информация в ряде случаев может выступать в качестве декларативной и наоборот.

3. Информатика и информационные системы

Под *информационной системой* понимают систему, организующую, хранящую и преобразующую информацию, то есть систему, основным предметом и продуктом в которой является информация.

По своей природе такие системы являются эрготехническими, в их функционировании принимают непосредственное участие и люди (*эргатические элементы*), и технические средства.

Информатика — это наука, изучающая свойства, структуру и функции информационных систем, основы их проектирования, создания, использования и оценки, а также информационные процессы, в них происходящие.

Информационные технологии – система процедур преобразования информации с целью ее формирования, организации, обработки, распространения и использования.

Индустрия информатики — это инфраструктурная отрасль хозяйства, обслуживающая другие отрасли материального производства и непроизводственной сферы, обеспечивающая их необходимыми информационными ресурсами, создающая условия для их эффективного функционирования и развития (своеобразная «нервная система» общественного производства).

4. Интеллектуальный анализ данных (ИАД)

Анализ данных (АД) – это система подходов и методов, ориентированная на выявление механизма порождения представленных данных в рамках имеющейся априорной модели этого механизма.

Современные технологии анализа данных – новая парадигма процесса исследования данных, основанная на принципах, предложенных Джоном Тьюки:

- Анализ – это способ существования данных. Его материальная основа – системы «человек – машина».
- Принцип многократного возвращения к одним и тем же данным.
- Принцип множественности возможных моделей.

- Принцип варьирования предпосылок с рассмотрением последствий такого варьирования.
- Принцип множественности результатов и выбора на основе неформальных процедур принятия решений.
- Принцип полного использования эндогенной информации и максимального учета информации экзогенной.

В ряде случаев АД строится и реализуется в соответствии с технологиями искусственного интеллекта.

NB! *Искусственный интеллект* (*ИИ*, artificial intelligence) — это общее понятие, описывающее «способность вычислительной машины моделировать процесс мышления за счет выполнения функций, которые обычно связывают с человеческим интеллектом»: построение и использование экспертных систем, логический вывод, понимание естественных языков, зрительное и слуховое восприятие (ГОСТ 15971 – 90. Системы обработки данных. Термины и определения).

NB! *Экспертная система* (ЭС, expert system) - это система искусственного интеллекта, включающая базу знаний с набором правил и *машину вывода* (inference engine), позволяющую на основании правил и предоставляемых пользователем фактов распознать ситуацию, сформулировать решение или дать рекомендацию. Обычно ЭС дополнительно включает в себя рабочий интерфейс пользователя, через который осуществляется взаимодействие эксперта с компьютером.

Таким образом, ЭС – это компьютерная система, которая эмулирует способности эксперта к принятию решения.

Объединение технологий АД и ИИ привело к возникновения нового направления обработки данных – интеллектуального анализа данных.

NB! *Интеллектуальный анализ данных (ИАД)* – исследование данных, использующее методы искусственного интеллекта и ориентированное на придание системе свойств искусственного интеллекта.

Вычислительная техника создавалась, прежде всего, для обработки данных. Рутинную часть анализа данных стараются переложить на *системы поддержки принятия решений (СППР, DSS)* – системы, обладающие средствами ввода, хранения и анализа данных из конкретной предметной области с целью поиска эффективного управляющего решения.

Такие системы не генерируют правильные решения, а предоставляют специалисту – аналитику данные в форме, удобной для изучения и анализа. *Интеллектуальные СППР* содержат функции, основанные на методах ИИ. Их главным отличием является способность к саморазвитию, проявляющаяся в генерации ка-

чественно новых решений, не предусмотренных исходными алгоритмами и программой.

1.5. Data Mining

Аналитик имеет дело и с документами, и с табличными значениями, которые также принято называть фактографическими.

Под единичным фактом принято понимать описание некоторого события. В формализованном виде для этого применяется следующая запись:

$$E_k = \{a_j, t, x_1, x_2, ..., x_m\}$$

где a_j – идентификатор (имя) объекта, t – время измерения, x_i – значение ***i***-й характеристики объекта.

NB! **Примечание.** Временной ряд событий образует временной ряд измерений. Существует два основных подхода к их формированию: во-первых, «по событию» – момент времени ***t*** определяется как момент изменения значения одной или более характеристик объекта; во-вторых, «по времени» – измерения проводятся через равные промежутки времени. Больший интерес с точки зрения практики представляет первый вариант, однако, большинство математических методов «заточено» под второй.

Рассматривая любой документ как множество высказываний можно гомоморфно отобразить его на множество фактов. Иначе говоря, из любого документа можно выделить некоторые факты. Именно они являются исходным сырьем для последующего анализа.

Обычно объекты предварительно упорядочиваются по некоторому признаку, как правило, представляющему собой одну из характеристик, которой обладают исследуемые объекты.

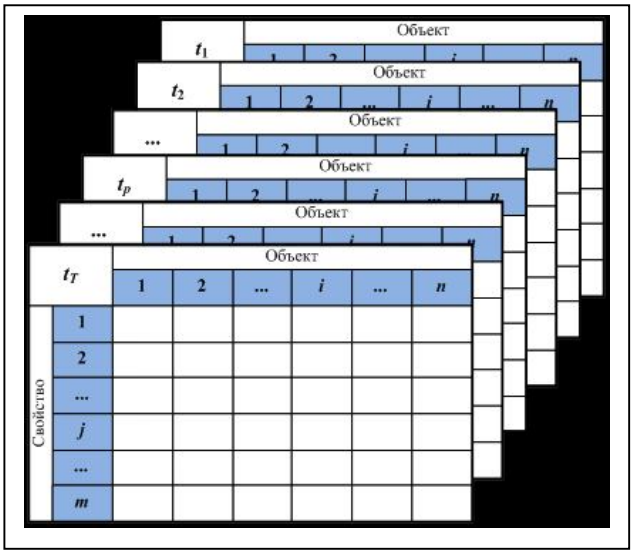


Рис. 1. Таблица «объект-свойство-время».

Фактографические данные, т.е. данные, непосредственно относящиеся к заданной предметной области, удобно представлять в табличном виде: строки a_1, a_2, …, a_n отражают информацию о самих исследуемых объектах (различаемых, как правило, по естественному (имени) или условному идентификатору), а столбцы x_1, x_2, …, x_m – информацию о значениях характеристик этих объектов. При необходимости учета временного фактора таких таблиц должно быть несколько: по одной на каждый отсчет времени.

Например, таблицы для медицинской диспансеризации содержат перечень студентов, (строки), а в столбцах стоят его индивидуальные параметры – пол, дата рождения, рост, вес, объем легких и т.п.

Машинная форма хранения данных содержит полезную информацию в скрытом виде, для ее извлечения и представления в удобном виде приходится использовать специальные методы. Технология *Data Mining* изучает именно процессы нахождения новых знаний в базах данных. В ее основе лежат

- Системы баз данных;
- Прикладная статистика;
- Теория искусственного интеллекта.

Data Mining переводится как "добыча" или "раскопка данных". Нередко рядом с Data Mining встречаются слова "обнаружение знаний в базах данных" (knowledge discovery in databases). Наиболее известная реализация технологий Data Mining – это поисковые системы в Интернете. В сфере бизнеса известны сообщения об экономическом эффекте от внедрения таких технологий, в 10-70 раз превысившем первоначальные затраты.

1.6. Закономерности

Выделяют пять типов закономерностей, которые позволяют выявлять методы Data Mining: *ассоциация, последовательность, классификация, кластеризация и прогнозирование.*

Ассоциация – это выделение различных типов связей между событиями: корреляционные связи, *if-then* правила и т.п.

Последовательность – это ассоциация между событиями, сдвинутыми во времени.

С помощью *классификации* выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил.

Кластеризация отличается от классификации тем, что сами группы заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных.

Основой для всевозможных систем *прогнозирования* служит историческая информация, хранящаяся в БД в виде временных рядов. Если удается построить найти шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем.

Примеры

1. В ходе расшифровки генома человека получены следующие выводы. Выделено примерно 35000 генов (17% объема), остальное – непонятные обломки. Для подавляющее большинства генов понятна их предыстория: такой ген был у рыб, у человека он развился таким-то образом. Отличие человека от шимпанзе –

около 350 генов, но из них 223 не имеют никакой предыстории, их происхождение непонятно. Очень хотелось бы так же изучить ДНК не из ядра, а из митохондрий, ее можно выделять из окаменевших костей, но это очень дорого. При этом можно было бы определить, когда эти новые гены попали в наследственность человека.

2. Холодная дождливая зима приводит к плохому урожаю и, одновременно, создает благоприятные условия для развития спорыньи – сорняка, содержащего наркотик ЛСД. Возникает цепочка: плохие урожаи – нарушение технологий выпечки хлеба – попадание в пищу ЛСД. В истории Европы обнаружена очень сильная корреляционная связь между этими событиями и непонятными психическими эпидемиями: плясками Святого Витта, вспышками бессмысленного насилия, массовыми сожжениями ведьм и колдунов. В частности, такие события наблюдались во Франции в 1793 году и в России – в 1917 и в 1928 г.г.

3. Анализ речей знаменитых ораторов – Троцкого, Гитлера, Фиделя Кастро и др., а также текстов, с помощью которых знахари заговаривают болезни, насылают и снимают порчу и т.п., привели к созданию специальной технологии - ***нейролингвистического программирования***, которую теперь широко используют в средствах массовой информации, в речах политиков, в рекламе и т.д.

Вопросы для самопроверки:

1. Назовите основную цель анализа данных?
2. Приведите наиболее распространенные определения информации.
3. Приведите формализованное описание факта.
4. Что называется фактографическими данными?
5. Перечислите основные составляющие экспертной системы?
6. Назовите основные задачи, решаемые средствами ИАД?
7. Перечислите принципы анализа данных, предложенные Дж. Тьюки?
8. Назовите три уровня анализа информации. Чем они отличаются?
9. Чем отличается классификация от кластеризации?
10. Назовите различие между данными и знаниями.

Литература:

1. Загоруйко Н.Г. Прикладной анализ данных и знаний. – Новосибирск : Изд-во НГУ, 1990.

2. Плэтт В. Информационная работа стратегической разведки: основные принципы. – М.: Изд-во иностр. лит-ры, 1958.

3. Паклин Н.Б., Орешков В.И., Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2013.

4. Барсегян А.А. и др. Анализ данных и процессов. – СПб: БХВ-Петербург, 2009.

5. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988.

6. Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977.

7. Орлов А.И. Прикладная статистика. – М.: Экзамен, 2004.

ЛЕКЦИЯ 2

АНАЛИТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

1.Аналитические информационные технологии в задачах управления

Эффективность функционирования любой системы в *существенной*, а в большинстве случаев, и *определяющей* степени зависит от эффективности используемой ей системы управления. При этом особую роль управление приобретает при работе со сложными динамическими объектами.

Под сложными объектами управления, в контексте изучаемого предмета, будем понимать многомерные многосвязные системы.

NB!

Заметим, что большинство реальных систем относятся именно к этой категории, вопрос лишь в степени приближения к реальности тех моделей, которые используются для их описания и изучения. Так, например, даже простейшие, так называемые, элементарные частицы, в физике, или простейшие организмы в биологии – амебы, вирусы, при внимательном изучении оказываются крайне сложными объектами, характеризуемым большим числом взаимодействующих элементов.

Для описания функционирования более сложных систем – человека, технологического или производственного процесса, корпорации, государства и т.п., практически всегда необходимо использовать многопараметрические модели, причем не только самих объектов, эволюционирующих во времени, но среды их взаимодействия.

Традиционная модель управления по Винеру (рис. 1) включает в себя системы наблюдения и собственно управления.

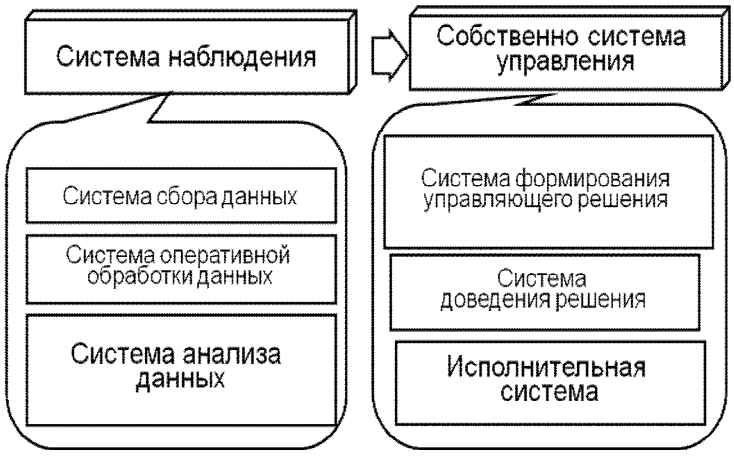


Рис. 1. Традиционная модель управления по Винеру

В последние годы особую роль приобретают вопросы автоматизации процесса выработки управляющих решений. На рис. 2 представлена та же общая схема управления с явно выделенной системой выработки управляющих решений. Автоматизированная система, ориентированная на задачу поддержки

принятия решений (или СППР) обычно используется в режиме «советника» или когнитивного ассистента.

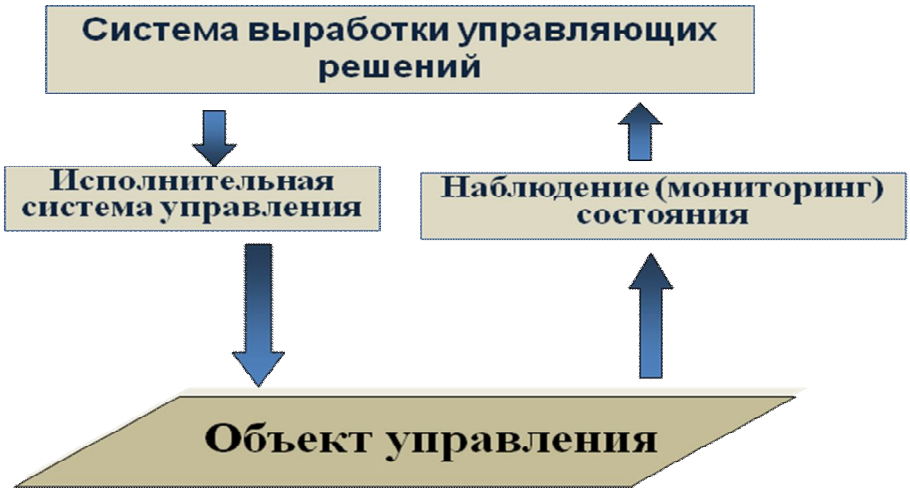


Рис. 2. Общая схема управления с явно выделенной системой выработки управляющих решений

Когнитивные информационные технологии ориентированы на создание качественно новых систем принятия решений, основанных на принципах искусственного интеллекта (artificial intelligence systems, AI). В настоящее время интеллектуальное управление строится, главным образом, на принципах гибридного интеллекта, когда в контуре управления ведущую роль исполняет человек, а компьютер формирует некоторые оптимизирующие рекомендации. В частности, ПЭВМ уже превратились в системы, восполняющие дефицит человеческой памяти, способности к арифметическим и логическим операциям, возможности по поиску и формированию знаний, необходимых для принятия решений.

В автоматизированных СУ формирование управляющих решений производится, как правило, *лицом, принимающим решение* (или ЛПР). В лучшем случае, в его распоряжении имеются средства OLTP – online treatment processing, а также простейшие схемы агрегации и визуализации данных.

Имеется ли реальная возможность формировать эффективные правленческие решения в указанных условиях?

Рассмотрим, в качестве примера, процесс управления реальным предприятием (рис. 3). Эффективность функционирования предприятия зависит от множества разнородных взаимосвязанных факторов. Эти факторы известны не полностью, некоторые из них являются латентными и проявляются лишь эпизодически. Как правило, отсутствуют точные представления о влиянии факторов на объем и качество производства, неизвестен характер динамических взаимозависимостей факторов между собой и между показателями качества выпускаемой продукции.

Рис. 3. Схема управления реальным предприятием

Вопрос: Может ли человеческий мозг справиться с задачей количественного анализа ситуации в условиях множества взаимосвязанных, неполностью известных факторов влияния?

Рис. 4. Развитие ситуации с 2-мя известными факторами влияния

Рассмотри простейший пример (рис.4) – прогноз развития линейного тренда под влиянием всего лишь двух факторов с минимальной нелинейностью (парабо-

лы 2-го порядка,) и точно известной функцией влияния - линейной функции с углом наклона 45^0. Попробуйте спрогнозировать динамику развития. Очевидно, что даже такая простейшая задача является для человека проблемой.

NB! *Вывод: Человеческий мозг в принципе не способен к эффективному количественному анализу и прогнозу развития ситуации.*

Это утверждение сохраняется и на других уровнях анализа данных – мировая динамика, экономика РФ, динамика развития предприятия, управление технологическими и производственными процессами, управление личной жизнью и т.п.

Разумеется, не следует и забывать о фантастической способности человека, несмотря ни на что, формировать в этих условиях, как правило, вполне сносные решения. Хотя и не всегда.

Мозг обладает некоторыми интегрирующими способностями восприятия и аналоговой переработки информации. Включаются механизмы правого полушария, ответственные за интуицию и т.п.

В связи с этим наилучшие результаты пока что достигаются при использовании человеко-машинного симбиоза, когда человек формирует стратегические решения, оценивает ситуацию в целом, верифицирует машинный результат и принимает окончательное решение, а ЭВМ – считает и работает в режиме «советчика» для количественного просчета ситуаций и подготовки (визуализации) данных.

NB! Слоган IBM: «Машина должна работать, человек – думать!»

В связи с этим базовая структура автоматического управления претерпевает определенные и существенные изменения, представленные на рис. 5.

В частности, базовая структура дополняется информационным хранилищем – базой производственного опыта и автоматизированной СППР.

Совокупность технических средств и методов, ориентированных на задачи автоматизированной поддержки принятия управленческих решений называется *аналитическими информационными технологиями (АИТ)*.

Основные направления АИТ включают в себя
- OTLP,
- OLAP,
- DW, Data Mart
- Data Mining.

Задачи хранения данных, их оперативной модификации, информационно-поискового анализа в условиях одновременного обращения многих пользователей решают системы *OLTP (Online Transactions Proceeding)*. Однако практика использования таких систем показала, что они плохо приспособлены к решению задач собственно анализа данных. Выход нашелся в создании специализированных подсистем - хранилищ данных [У. Инмон, 1992].

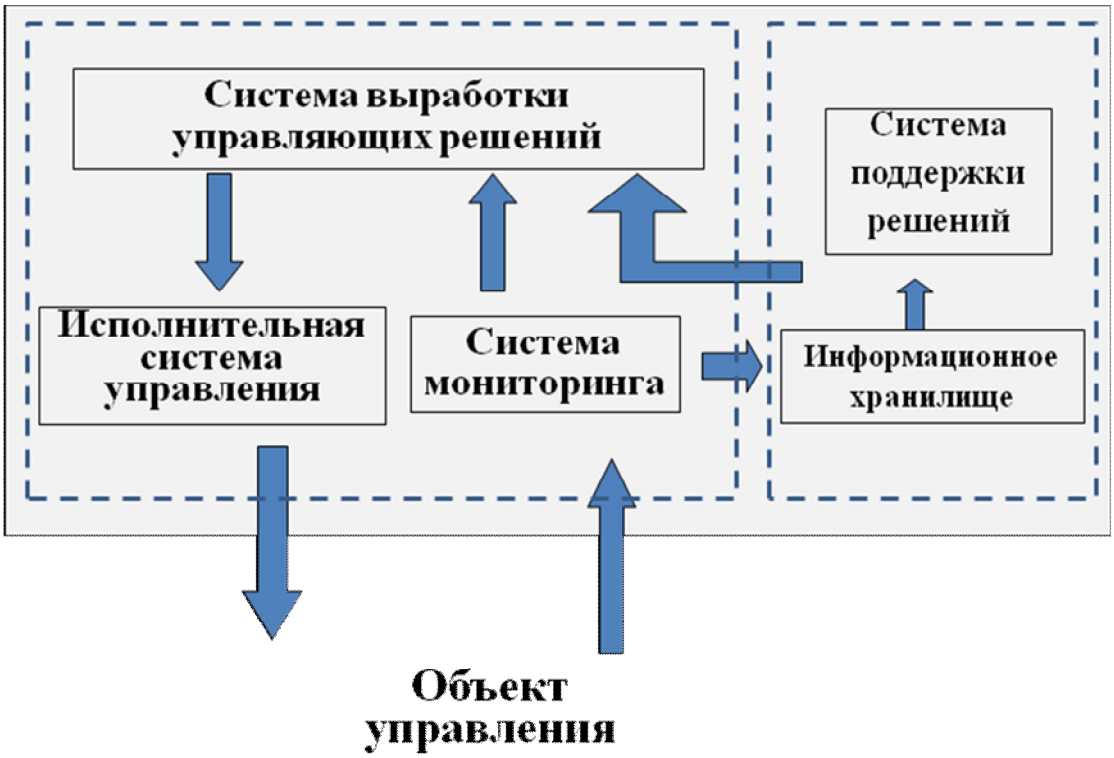


Рис. 5. Схема когнитивного управления

Хранилище данных (ХД, Data Warehouse) – предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений. ХД может быть как физическим, так и виртуальным. Обычно данные для ХД копируются критически, очищаются и обогащаются новыми атрибутами.

Витрина данных (ВД, Data Mart) – упрощенный вариант ХД, содержащий только тематически объединенные данные. ВД часто формируют как надстройки над более общим ХД.

В 1993г. Е. Кодд, основоположник реляционной модели БД, предложил представление данных в виде многомерной модели гиперкуба, ребрами которого являются измерения или параметры изучаемого объекта. Эту технологию назвали *OLAP (Online analytical processing)*, ее полное определение задается 12 правилами Кодда, приведенными в Приложении 1.

Набор этих требований, послуживших де-факто определением OLAP, достаточно часто вызывает различные нарекания, например, правила 1, 2, 3, 6 являются требованиями, а правила 10, 11 — неформализованными пожеланиями. Таким образом, перечисленные 12 требований Кодда не позволяют точно определить OLAP. В 1995 г. Кодд добавил еще шесть правил:

13. Пакетное извлечение против интерпретации — OLAP-система должна в равной степени эффективно обеспечивать доступ как к собственным, так и к внешним данным.

14. Поддержка всех моделей OLAP-анализа — OLAP-система должна поддерживать все четыре модели анализа данных, определенные Коддом: толковательную, стереотипную, категориальную и умозрительную.

15. Обработка ненормализованных данных — OLAP-система должна быть интегрирована с ненормализованными источниками данных. Модификации данных, выполненные в среде OLAP, не должны приводить к изменениям данных, хранимых в исходных внешних системах.

16. Сохранение результатов OLAP: хранение их отдельно от исходных данных — OLAP-система, работающая в режиме чтения-записи, после модификации исходных данных должна сохранять результаты отдельно друг от друга, т.е. обеспечивать безопасность всех исходных данных.

17. Исключение отсутствующих значений — OLAP-система, представляя данные пользователю, должна отбрасывать все отсутствующие значения, т.е. они должны отличаться от нулевых значений.

18. Обработка отсутствующих значений— OLAP-система должна игнорировать все отсутствующие значения без учета их источника. Эта особенность связана с 17-м правилом.

Data Mining – исследование и обнаружение машиной (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком

Вернемся к предприятию или любому иному объекту управления. Эффективно управлять – значит научиться достоверно предсказывать, к чему приведут последствия от реализации принятого управленческого решения.

Исполнение отстает от решения, эффект почти всегда задержан во времени по отношению к исполнению, следовательно, эффективное управление практически всегда носит прогностический характер.

Прогноз предназначен для оценки состояния системы в будущем. При этом предполагается, что система является динамической, т.е. развивающейся под воздействием множества факторов влияния – собственных и факторов среды. Отсюда для достоверного количественного прогноза необходимо научиться оперативно обнаруживать факторы влияния, скрытые зависимости и т.п.

Вопрос: Что мешает формировать высокоэффективный прогноз? Прежде всего, неполнота информации и ее недостаточная достоверность.

Основные проблемы эффективного управления включают в себя решение следующих задач:

- Выявление значимых факторов влияния;
- Определение взаимосвязей факторов;
- Определение тенденций развития;
- Прогнозирование результатов;
- Оптимизацию решений.

Решение перечисленных проблем составляет центральную задачу ИАД (DM) и поддерживающих ее АИТ.

По постановке задачи разделяют на ***обучение с учителем*** (*Supervised Learning*) и ***обучение без учителя*** (*Unsupervised Learning*). Для управления полученными в результате анализа знаниями используются технологии ***Knowledge Management***.

2. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

DM не имеет одного отца-основателя, его создавали тысячи математиков-прикладников, работающих в области компьютерной обработки данных.

Сфера применения Data Mining ничем не ограничена – она везде, где имеются какие-либо данные. Но в первую очередь методы Data Mining сегодня, мягко

говоря, заинтриговали коммерческие предприятия, развертывающие системы хранения больших данных (Big Data) на основе информационных хранилищ данных (Data Warehouse).

Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигнуть 1000% . Например, известны сообщения об экономическом эффекте, в 10-70 раз превысившем первоначальные затраты от 350$ до 750$ тыс. Известны сведения о проекте в $ 20 млн., который окупился всего за 4 месяца. Другой пример – годовая экономия $700 тыс. за счет внедрения Data Mining в сети универсамов в Великобритании.

Сейчас в мире действует множество фирм, занятых в индустрии производства продуктов DM, включаю такие гиганты, как Microsoft, Oracle, SAS Institute и др. В последние годы за рубежом появилось множество монографий и учебных пособий в данной области.

DM представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов DM они могут получать ощутимые преимущества в конкурентной борьбе.

Одно из возможных определений DM:

Data Mining (или интеллектуальный анализ данных) – направление в области информационных и математических технологий, направленное на решение задач анализа данных в интересах повышения эффективности управляющих решений.

Соответственно, назначение DM состоит в решении задач в интересах систем поддержки принятия решений на основе количественных и качественных исследований сверхбольших массивов разнородных ретроспективных данных.

Основные задачи, решаемые средствами DM, представлены на рис. 6

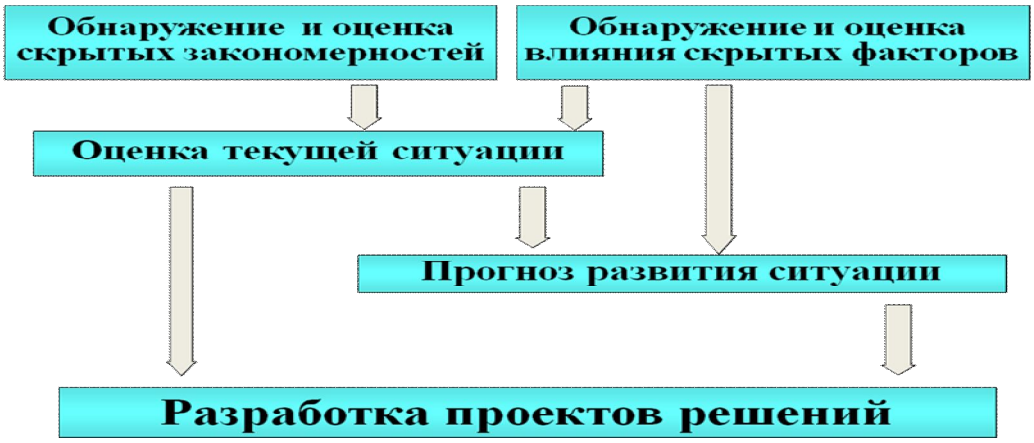


Рис. 6. Основные задачи, решаемые средствами DM

Как видно из приведенного на рис. 6 списка, перечень базовых задач, решаемых средствами DM, полностью совпадает со список задач, с которым сталкивается менеджер при управлении практически любым предприятием или объектом - ТП, финансами, коммерческой деятельностью и т.п. А именно - поиск закономерностей, взаимосвязей, факторов влияния, угроз, прогнозирование и поиск возможных решений.

Математический инструментарий DM представлен на рис. 7.

Как видно из представленной картинки, математический арсенал DM включает в себя почти все направления современной прикладной математики. Однако особое внимание уделяется статистическим методам обработки, обеспечивающим возможность использовать накопленный статистический опыт управления предприятием, и новейшие кибернетические методы, среди которых особенно следует отметить нейросетевые технологии, генетические алгоритмы, методы эволюционного программирования и др.

Важной особенностью математического инструментария DM является его реализация в виде законченных программных продуктов, как правило, коммерческих.

Рис. 7. Математический инструментарий DM

В частности, в табл. 1 приведены некоторые программные продукты, относящиеся к категории DM, и их ориентировочные стоимости.

3. СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЯ

Для того, чтобы рассмотреть возможность применения DM-технологий в СППР рассмотрим первоначально *традиционную структуру* такой системы. Из схемы на рис. 8 видно, что традиционная СППР ориентирована на оперативную работу эксперта и включает в себя различные средства обеспечения OLTP анализа.

Развитие СППР на основе DM предполагает модификацию СППР, как это показано на рис. 9.

Как видно из приведенной структуры, аналитическая СППР дополнительно включает в себя информационное хранилище и АРМ эксперта-аналитика, работающего с инструментами DM.

Таблица 1. Программные продукты, относящиеся к категории DM

DM классы	**Системы**	**Стоимость**
Предметно-ориентированные аналитические системы	Скрининговые системы, ИС ЛПУ, ИС врача, ИС фельдшера, инф-справ. ИС и др	$ 300-20000
Статистический анализ	SPSS, SAS, STATGRAPHICS, STATISTICA, STADIA	$1000-15000
Нейронные сети	BrainMarker, NeuroShell, OWL	$ 1500-8000
Ассоциации по аналогии	CBR, KATE Tools, Pattern Recognition Workbench	$1500 -10000
Деревья решений	See5/C5.0, Clementine, SIPINA, KnowledgeSEEKER	$1000 -10000
Эволюционное программирование	PolyAnalyst, NeuroShell	$1000 -5000
Генетические алгоритмы	GeneHunter	$1000
Алгоритмы ограниченного перебора	WizWhy	$4000
Системы визуализации многомерных данных	DataMiner3D	До $1000

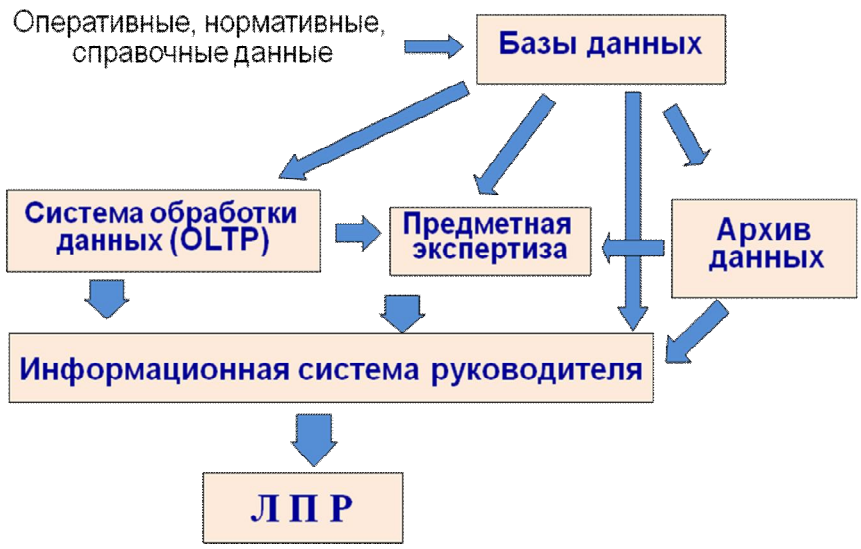


Рис. 8. Структура традиционной СППР

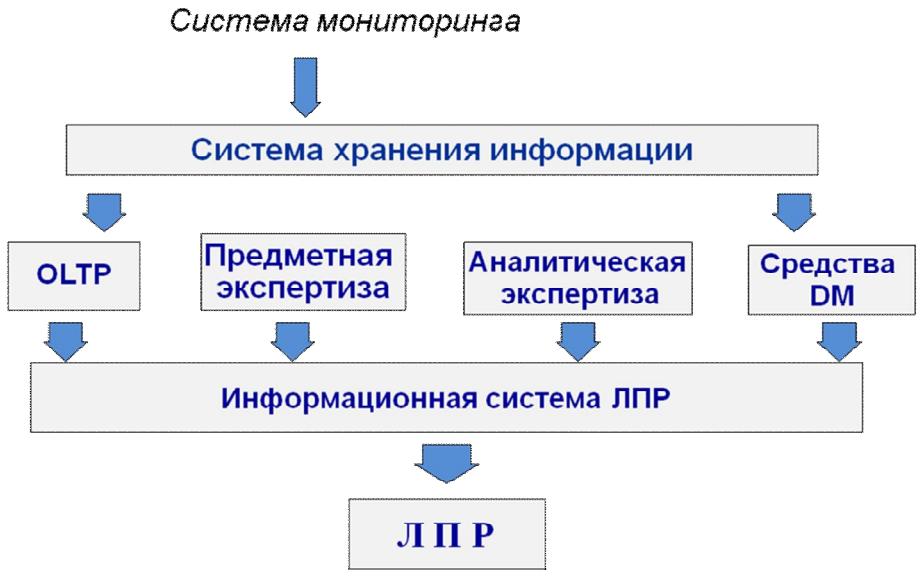


Рис. 9. Структура аналитической СППР

Особое место в АИС занимают **системы хранения информации** (рис. 10).

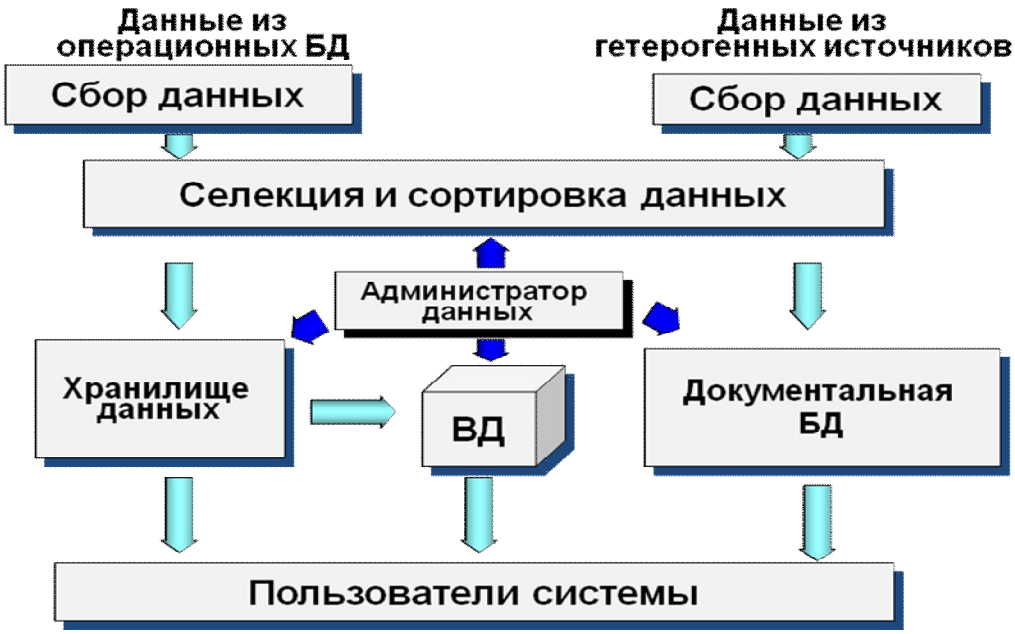


Рис. 10. Структура системы хранения информации для аналитической СППР

Предполагается, что вместо традиционной реляционной БД будут использоваться БД с многомерным представлением данных, образующих *информационное хранилище (DW, Data Warehouse)*. Локальная многомерная БД, как правило, имеющая тематическую ориентацию, получила название *«витрина данных»* или *Data Mart***.**

Основные принципы построения хранилищ данных были сформулированы Б. Инмоном и Коддом в концепции *OLAP*.

Существенная проблема, приведшая к многомерным структурам хранения данных, – скорость обработки аналитических запросов.

Важный принцип построения DW – его неизменность по отношению хранящимся в нем историческим данным.

4. Data Mining, как средство добычи знаний

Отметим два основных концептуальных направления в разработке алгоритмов DM (рис. 12).

Первое направление связано с использованием усредненного опыта. К этому направлению можно отнести все статистические алгоритмы анализа данных.

По существу, технология формирования и накопления устойчивых представлений об окружающем мире так или иначе связана с усредненным опытом.

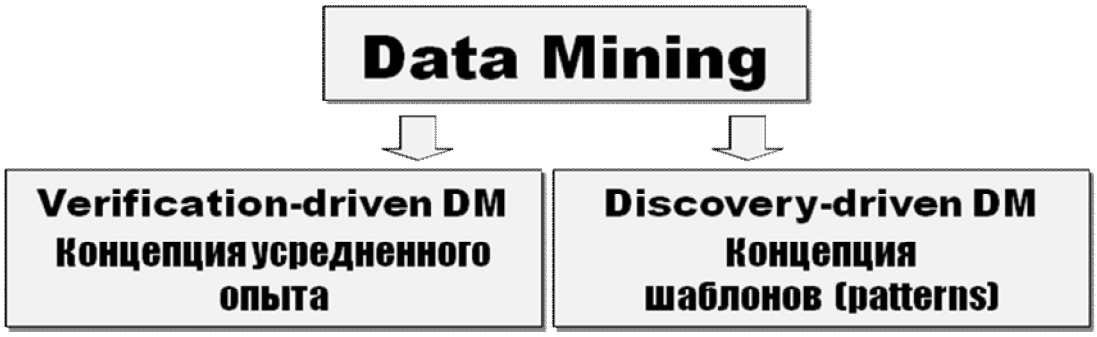


Рис. 12. Основные направления Data Mining

Однако в ряде случаев приходиться общаться и с уникальными ситуациями. Более того, каждая конкретная ситуация является уникальной и неповторимой. Согласно демокритовской концепции диалектической изменчивости, «все течет, все меняется». В этом случае используется концепция шаблонов или паттернов (patterns).

Для того, чтобы понять или осмыслить текущую необходимо сопоставить ее с базой знаний усредненного опыта, которая хранится в памяти человека или в БД компьютера. При этом возникает необходимость в формировании меры «похожести» или «близости», позволяющей для каждой ситуации выбрать из БД наиболее подходящий шаблон, позволяющий интерпретировать полученные наблюдения.

Современные подходы к задачам управления сложными динамическими системами выдвигают ряд новых требований, часто являющиеся достаточно противоречивыми. Важнейшими их них являются:

- сверхбольшой объем данных;
- разнородность данных;
- глубина анализа;
- интерпретируемость данных;
- доступность (простота) инструментария.

Сверхбольшой объем данных связан с проблемой больших данных Big Data. Человечество накопило огромные массивы цифровых данных, однако они ничего не дают без соответствующих технологий извлечения из них полезных знаний.

Разнородность данных, связанное с большим количеством плохо структурированной информации (тексты, -аудио, -видео, рисунки и т.п.) привела к проблеме Data Fusion – слияния данных и приведение их к формам, доступным для автоматического анализа данных.

Интерпретируемость данных – требование, связанное с представлением данных в форме, доступной для человеческого мозга. так например, уже 4-х мерных данные не допускают наглядных геометрических представлений.

Доступность инструментария предполагает, что предложенными математическими моделями и технологиями могут пользоваться предметные эксперты и специалисты, не имеющие специальной математической подготовки.

Глубина анализа данных может осуществляться на трех уровнях, представленных на рис. 13. Классификатор глубины анализа включает в себя поверхностный, неглубокий и глубокий уровни анализа, которым отвечают, соответственно, OLTP-, OLAP и DM технологии.

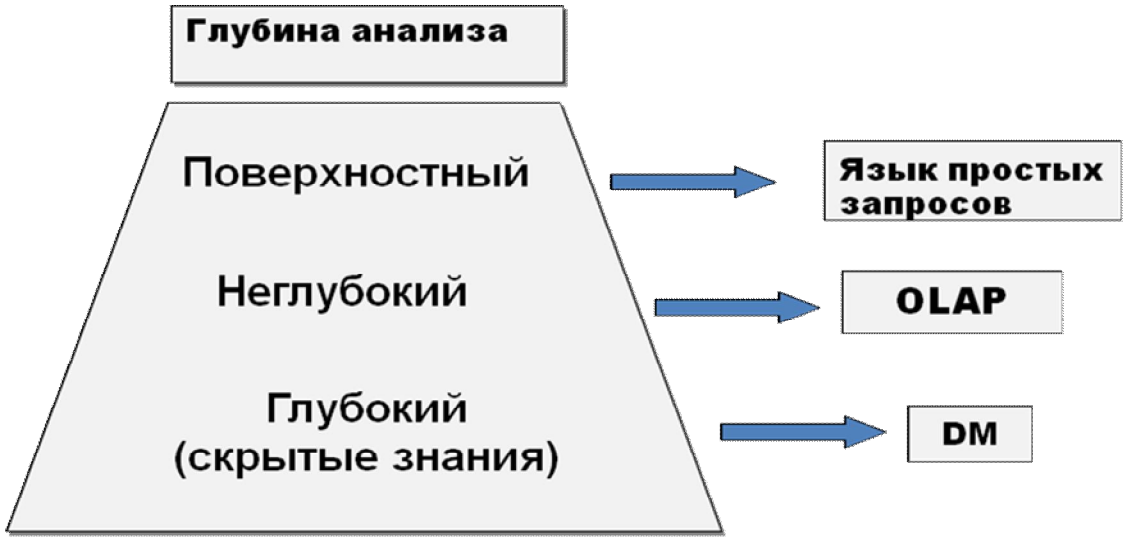


Рис. 13. Три уровня глубины анализа

Глубокий анализ данных включает в себя следующие основные задачи:
- ассоциация – выявление связей между событиями;
- обнаружение последовательностей;
- выявление связанных во времени событий;
- классификация – выявление признаков групп событий;
- кластеризация – тоже для заранее не выявленных групп событий;
- прогнозирование.

Сравнение OLAP и DM технологий представлено в таблице 2.

Реализация процедуры анализа, в общем случае, представляет собой сложную многоступенчатую процедуру. По мнению западных методистов, основные шаги к успеху анализа данных включают в себя:

1. Четкое представление о цели;
2. Сбор релевантных данных;
3. Выбор методов анализа;
4. Выбор программного средства;
5. Выполнение анализа;
6. Принятие решения.

Таблица 2. Сравнение OLAP и DM технологий

OLAP	DM
- Средние показатели реализаций под влиянием различных факторов	- Наличие аналогов и подобий в предыстории изучаемого события
- Девиации, максимальные и минимальные отклонения от нормы;	- Наличие и характер взаимозависимостей между наборами событий;
- Графические представления динамики агрегированных данных.	- Прогнозирование эволюции развития ситуаций под влиянием набора взаимосвязанных факторов

В заключение лекции приведем пример перечня задач, решаемых средствами DM, в интересах промышленного производства:

- выявление скрытых закономерностей и взаимосвязей в динамики состояния технологических процессов (ТП);
- выявление наиболее значимых факторов влияния на качество функционирования ТП;
- оценка вероятности выхода показателей качества выпускаемой продукции и значений параметров состояния ТП за допустимые пределы;
- прогнозирование изменения показателей качества и объемов выпуска товарной продукции в зависимости от выбора стратегии и режимов управления;
- формирования оптимальных вариантов управления ТП.

Вопросы для самопроверки:

1. Как устроена общая (кибернетическая) модель управления?
2. Приведите перечень задач, связанных с управлением реальным производственным предприятием.
3. Чем отличаются дескриптивные задачи от предсказательных?
4. Что называется сложными объектами?
5. Перечислите основные направления анализа данных, входящие в АИТ?
6. Назовите основные задачи, решаемые средствами ИАД?
7. Перечислите проблемы построения эффективного прогноза?
8. Назовите особенности хранения информации в аналитических системах.
9. Чем отличается когнитивное управление от традиционного?
10. Перечислите математический инструментарий Data Mining.

Литература:

1. Загоруйко Н.Г. Прикладной анализ данных и знаний. – Новосибирск : Изд-во НГУ, 1990.
2. Барсегян А.А. и др. Анализ данных и процессов. – СПб: БХВ-Петербург, 2009.
3. Дюк В., Самойленко А. Data Mining: учебный курс (+CD). — СПб.: Изд. Питер, 2001. — 368 с.

4. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. — М.: Изд. «Фазис», 2006. — 176 с. .

5. Зиновьев А. Ю. Визуализация многомерных данных. — Красноярск: Изд. Красноярского государственного технического университета, 2000. — 180 с.

6. Чубукова И. А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. — 382 с.

7. Ian H. Witten, Eibe Frank and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664.

Приложение 1. Правила Кодда для OLAP систем

В 1993 году Кодд опубликовал труд под названием "*OLAP* для пользователей-аналитиков: каким он должен быть". В нем он изложил основные концепции оперативной аналитической обработки и определил 12 правил, которым должны удовлетворять продукты, предоставляющие возможность выполнения оперативной аналитической обработки.

1. Концептуальное многомерное представление. *OLAP*-модель должна быть многомерной в своей основе. Многомерная концептуальная схема или пользовательское представление облегчают моделирование и анализ так же, впрочем, как и *вычисления*.

2. Прозрачность. Пользователь способен получить все необходимые данные из *OLAP*-машины, даже не подозревая, откуда они берутся. Вне зависимости от того, является *OLAP*-продукт частью средств пользователя или нет, этот факт должен быть незаметен для пользователя. Если *OLAP* предоставляется *клиент*-серверными вычислениями, то этот факт также, по возможности, должен быть невидим для пользователя. *OLAP* должен предоставляться в *контексте истинно* открытой архитектуры, позволяя пользователю, где бы он ни находился, связываться при помощи аналитического инструмента с сервером. В дополнение к этому прозрачность должна достигаться и при взаимодействии аналитического инструмента с гомогенной и *гетерогенной* средами *БД*.

3. Доступность. *OLAP* должен предоставлять свою собственную *логическую схему* для доступа в *гетерогенной* среде *БД* и выполнять соответствующие преобразования для предоставления данных пользователю. Более того, необходимо заранее позаботиться о том, где и как, и какие типы физической организации данных действительно будут использоваться. *OLAP*-система должна выполнять доступ только к действительно требующимся данным, а не применять общий принцип "кухонной воронки", который влечет ненужный ввод.

4. Постоянная производительность при разработке отчетов. *Производительность* формирования *отчетов* не должна существенно падать с ростом количества измерений и размеров базы данных.

5. *Клиент*-серверная архитектура. Требуется, чтобы продукт был не только *клиент*-серверным, но и чтобы серверный компонент был бы достаточно интеллектуальным для того, чтобы различные клиенты могли подключаться с *минимумом* усилий и программирования.

6. Общая многомерность. Все измерения должны быть равноправны, каждое измерение должно быть *эквивалентном* в структуре, и в операционных возможностях. Правда, допускаются дополнительные операционные возможности для отдельных измерений (видимо, подразумевается время), но такие дополнительные функции должны быть предоставлены любому измерению. Не должно быть так, чтобы базовые *структуры данных*, вычислительные или отчетные форматы были более свойственны какому-то одному измерению.

7. Динамическое управление *разреженными матрицами*. *OLAP* системы должны автоматически настраивать свою физическую схему в зависимости от *типа модели*, объемов данных и разреженности базы данных.

8. Многопользовательская *поддержка*. *OLAP*-инструмент должен предоставлять возможности *совместного доступа*(запроса и дополнения), *целостности* и безопасности.

9. Неограниченные перекрестные операции. Все виды операций должны быть дозволены для любых измерений.

10. Интуитивная манипуляция данными. Манипулирование данными осуществлялось посредством прямых действий над ячейками в режиме просмотра без использования меню и множественных операций.

11. Гибкие возможности получения *отчетов*. Измерения должны быть размещены в отчете так, как это нужно пользователю.

12. Неограниченная *размерность* и число уровней *агрегации*. Исследование о возможном числе необходимых измерений, требующихся в аналитической модели, показало, что одновременно может использоваться до 19 измерений. Отсюда вытекает настоятельная рекомендация, чтобы аналитический инструмент был способен одновременно предоставить как *минимум* 15 измерений, а предпочтительнее 20. Более того, каждое из общих измерений не должно быть ограничено по числу определяемых пользователем-аналитиком уровней *агрегации* и путей *консолидации*.

ЛЕКЦИЯ 3

МИР МОДЕЛЕЙ и АНАЛИЗ ДАННЫХ. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ РЕАЛЬНОСТИ

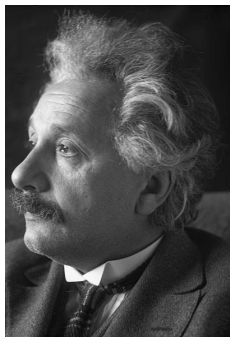
1. Введение. Математика – язык научных исследований

Основой анализа данных является получение знаний. В некоторых случаях, например, при проведении академических исследования, эти знания должны обладать высоким уровнем общности, т.е. относиться к категории фундаментальных. В свою очередь «фундаментальность» означает самый высокий, математический уровень их обоснованности. И здесь "во главе угла" встает Ее Величество Математика и ее прикладные ответвления – математическая статистика, теоретическая кибернетика, теория систем, исследование операций и т.п.

Здесь уместно вспомнить высказывание А. Эйнштейна: *"Чисто математические построения позволяют найти те понятия и те закономерные связи между ними, которые дают ключ к пониманию явлений природы".*

Альберт Эйнштейн (1879, Ульм, Вюртемберг, Германия —1955, Принстон, Нью-Джерси, США) — физик-теоретик, один из основателей современной теоретической физики, лауреат Нобелевской премии по физике 1921 года, общественный деятель-гуманист.

Жил в Германии (1879—1893, 1914—1933), Швейцарии (1893—1914) и США (1933—1955). Почётный доктор около 20 ведущих университетов мира, член многих Академий наук, в том числе иностранный почётный член АН СССР (1926). Эйнштейн — автор более 300 научных работ по физике, а также около 150 книг и статей в области истории и философии науки, публицистики и др.



В мире царят закономерности, вполне описываемые математическими уравнениями. И люди настолько привыкли к этим закономерностям, что просто не хотят ни замечать их, ни удивляться самому факту их возникновения.

Так, например, философские *законы диалектики* или *законы эволюции живой природы и человеческой цивилизации* представляют собой настоящее чудо, поскольку их генезис (как и происхождение большинства других окружающих нас законов) по сей день остается неясным. Даже для наиболее глубоко изученных физических закономерностей первопричины гравитации, света, электрического тока и других явлений остаются нераскрытыми.

Известный американский математик профессор нью-йоркского университета М. Клайн писал: *"Если бы Платон... писал Библию, он, несомненно, начал бы ее такими словами: "Вначале Бог создал математику, а затем небо и звезды согласно законам математики*".

Однако и математика не всесильна. Математика в лучшем случае описывает структуру закономерностей, но отнюдь не гарантирует их понимание.

Как писал Б. Рассел, *"математика представляет собой собрание выводов, которые могут быть применены к чему угодно".*

Семантика и интерпретация найденных физических и иных законов требуют их осознания на более высоком уровне - на уровне причин их происхождения и места во всеобщей картине мироздания. Строго говоря, знания современных ученых не так уж далеко ушли вперед в понимании первопричин фундаментальных законов по сравнению со знаниями средневековых схоластов, ссылающихся на Божественный промысел.

Математика - наука очищения, она позволяет обнаружить и идентифицировать закономерности, замаскированные в зашумленных стохастических потоках энергии и информации. Как правило, глубокие, фундаментальные закономерности достаточно просты и допускают красивые аналитические описания. *"Наш опыт убеждает нас,* - писал А. Эйнштейн, - *что природа - это реализация самых простых математических идей".* А математические описания реальных закономерностей и связей и представляют собой класс математических моделей.

Элегантность основополагающих законов оказалась вполне адекватной красоте классической "чистой" математики. Возможно, именно этот факт привел Г. Харди к мысли о том, что *"красота есть пробный камень для математической идеи; в мире нет места уродливой математике".*

Увы... Стоит только спуститься со снежных вершин ортодоксальной Науки к решению прикладных задач суетного Повседневья, и мы сразу же столкнемся с целым рядом проблем, явно не вписывающихся в изящные формы классических математических моделей. И здесь Ее Величеству Чистой Математике приходится сбросить с себя белоснежную мантию теоретической аналитики, засучить рукава и, превратившись в работящую служанку - Прикладную Математику, заняться тяжелым вычислительным трудом - имитационным моделированием, приближенными методами расчета, линеаризацией, обработкой некондиционных статистических данных и т.д. и т.п.

Следует иметь в виду, что *"математика - не свод готовых ответов на любой вопрос. Математика - это скорее школа мышления"* [15]. И важнейшим атрибутом такого мышления является системность, т.е. способность увидеть взаимосвязанное единство в совокупности разрозненных фактов, событий, элементов.

2. Моделирование. Понятие модели. Классификация

Главным инструментом изучения любых систем и связанных с ними процессов является ***моделирование***.

1 Под термином *модель* (от лат. *modulus* — «мера, аналог, образец») обычно понимают некоторое упрощенное представление какого-то реального объекта или протекающих в нем процессов.

Построение и исследование моделей, то есть *моделирование*, предназначено для изучения свойств и закономерностей, имеющихся в окружающем человека мире.

Существует множество различных видов моделирования. Например:

- Информационное моделирование
- Компьютерное моделирование
- Математическое моделирование
- Математико-картографическое моделирование
- Молекулярное моделирование

- Цифровое моделирование
- Логическое моделирование
- Педагогическое моделирование
- Психологическое моделирование
- Статистическое моделирование
- Структурное моделирование
- Физическое моделирование
- Экономико-математическое моделирование
- Имитационное моделирование
- Эволюционное моделирование
- Графическое и геометрическое моделирование и т. д.

По существу, весь процесс адаптации человека к окружающей его действительности основан на перманентном *ментальном моделировании*.

Процесс осознания информации, поступающей в мозг человека от сенсорных органов восприятия (зрение, слух, осязание et cetera), осуществляется путем естественного *подсознательного формирования ментальных моделей*, которые de facto интерпретируются как отображения реальной действительности. На самом деле этого всего лишь модели, формируемые мозгом и отображающие реальность весьма приближенно и неоднозначно.

В некоторых случаях ментальные модели формируют заведомо некорректную информацию об окружающем мире. В то же время, в некоторых случаях подобные деформации позволяют человеку более остро воспринимать свойства реального мира, важные для его выживания и эволюции. Так, например, цветовая гамма визуальных отображений представляет собой иллюзию, формируемую мозгом и позволяющую на сенсорном уровне различать электромагнитное излучение оптического диапазона для различных участков его частотного спектра. Если бы мир состоял только из дальтоников, то само понятие «цвета» в принципе было бы недоступно для понимания.

По сути, человек живет в виртуальном мире ментальных моделей, аналогичных миру теней в платоновской пещере. Даже представление о времени является виртуальным, ведь прошлого уже нет, будущее еще не наступило, а настоящее – бесконечно малый промежуток времени между прошлым и будущим, не доступный для восприятия в силу своей краткосрочности. Тем не менее, этот виртуальный мир ментальных моделей позволяет ориентироваться, выживать и даже преуспевать в объективно существующим реальном мире (разумеется, при выполнении непроверяемой гипотезы о его существовании).

Осознание виртуального характера восприятия реальности приводит к весьма непростым «философским» вопросам:

- если de facto мы живем в мире ментальных моделей, то какова вероятность того, что мы вообще существуем в реальном мире (или в реале), а не в некотором аналоге информационной «Матрицы»? Чем виртуальная жизнь отличается от реальной с позиции ее восприятия на уровне ментального моделирования?

- для чего нужен реальный мир? Если для энергетического поддержания человеческого организма, как формирователя и носителя ментальных отображений, то не легче ли перейти к более энергетически эффективным носителям, например, полупроводниковым? Если реал используется для процесса самоутверждения личности и достижения материального превосходства над себе подобными (с использованием таких социальных категорий, как власть, деньги, извест-

ность и т.п.), то не легче этого добиться в виртуале на основе программы виртуальной реальности, где каждый может построить желательный сценарий жизненного цикла (и не один). Например, любой мужчина может выбрать программу миллиардера, супергероя-спасителя человечества или абсолютного властелина мира.

- является ли человечество венцом эволюции, либо мы, как продукт химико-биологической эволюции, являемся лишь промежуточным этапом, некоторой проходящей формой к значительно более мощной в интеллектуальном смысле кристаллической форме разумной жизни?

Вернемся к процессу моделирования, как к инструменту познания объективного мира. Как правило, изучение реального мира опирается на предположение о том, что сложность любого материального объекта и окружающего его мира бесконечна вследствие неисчерпаемости материи и форм её взаимодействия внутри себя и с внешней средой. Для изучения материального объекта или процесса нужна не одна, а целая *система моделей,* отображающих его различные аспекты и свойства. Как следствие, существует много типов моделей, большинство из которых отражает решение некоторой конкретной задачи. Таким образом, наиболее общая форма представления знаний связана с концепцией мультимодельности.

Рассмотрим простейшую классификацию наиболее общих видов моделей. В частности, *по способу отображения действительности* различают три основных вида моделей:

- *эвристические,*
- *натурные и полунатурные;*
- *математические*.

Эвристические модели – это частный вид *ментальных моделей*, связанных с представлением образов в воображении человека. Обычно такие модели имеют вид дескриптивных описаний в терминах естественного языка (*вербальная информационная модель*). Очевидно, что такие модели являются заведомо субъективными, неоднозначными и некорректными. «Сколько людей, столько мнений» - «Quot homines, tot sententiae». (Первоисточник — сочинения римского драматурга Теренция, Публий Теренций Афр, ок. 195—159 до н.э.).

Эвристические модели не описываются формально-логическими и математическими выражениями, однако они являются предтечами более строгих, формализованных моделей. Этап эвристического или ментального моделирования является принципиально необходимой частью познания реальных процессов и явлений.

Эвристическое моделирование осуществляется за счет креативности сознания и являются основным средством вырваться за пределы устоявшихся знаний. Но способность к такому моделированию зависит, прежде всего, от творческой фантазии человека, его опыта и эрудиции.

Эвристические модели используют на начальных этапах проектирования или других видов деятельности, когда сведения о разрабатываемой системе ещё скудны. На последующих этапах проектирования эти модели заменяют на более конкретные и точные технологии моделирования.

Натурное моделирование основано на формирова-

нии моделей, подобных реальным объектам. Отличие натуральных моделей от реальных объектов и процессов состоит в их размерах, численности элементов, материале элементов и т. п. В некоторых случаях используется *полунатурное моделирование*, когда часть объекта исследований заменяется имитационной моделью.

В системах управления и информационных технологиях крайне важное значение имеют формализованные м*атематические модели*.

Математические модели представляют собой совокупность взаимосвязанных математических, формально-логических выражений, как правило, отображающих реальные процессы и явления (физические, психические, социальные и т. д.).

В теоретических исследованиях используются математические модели, в экспериментальных - натурные, полунатурные и т.п. Задача исследователя состоит в том, чтобы, применяя системную методологию, математическое или иное моделирование и другой арсенал научных исследований, решить стоящую перед ним научную проблему.

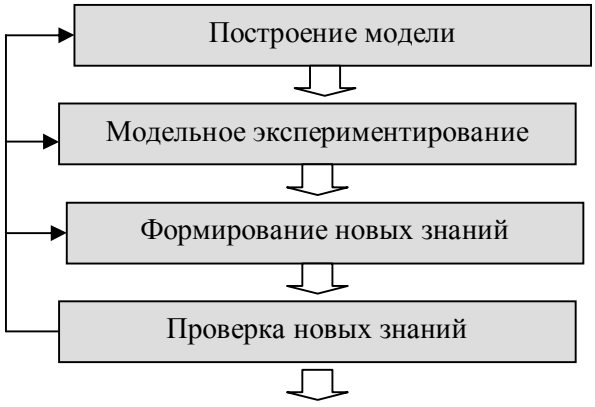


Рис. 1. Основные этапы моделирования

Сам процесс моделирования включает в себя ряд этапов, представленных на рис. 1.

Первый этап построения модели предполагает наличие некоторых знаний об объекте-оригинале. Познавательные возможности модели обусловливаются тем, что модель отображает (воспроизводит, имитирует) какие-либо существенные черты объекта-оригинала.

Вопрос о необходимой и достаточной мере сходства оригинала и модели требует конкретного анализа. Очевидно, модель утрачивает свой смысл как в случае тождества с оригиналом (тогда она перестает быть моделью), так и в случае чрезмерного во всех существенных отношениях отличия от оригинала. Таким образом, изучение одних сторон моделируемого объекта осуществляется ценой отказа от исследования других сторон. Поэтому любая модель замещает оригинал лишь в строго ограниченном смысле.

Из этого следует, что для одного объекта может быть построено несколько «специализированных» моделей, концентрирующих внимание на определенных сторонах исследуемого объекта или же характеризующих объект с разной степенью детализации.

На втором этапе модель выступает как самостоятельный объект исследования. Одной из форм такого исследования является проведение «модельных» экспериментов, при которых сознательно изменяются параметры или условия функционирования модели и систематизируются данные о ее «поведении». Конечным результатом этого этапа является совокупность знаний о модели, о ее поведении в различных условиях.

На третьем этапе осуществляется перенос знаний с модели на оригинал — формирование новых знаний. Одновременно происходит переход с «языка» мо-

дели на «язык» оригинала. Процесс переноса знаний проводится по определенным правилам. Знания о модели должны быть скорректированы с учетом тех свойств объекта-оригинала, которые не нашли отражения или были изменены при построении модели.

Четвертый этап — практическая проверка получаемых с помощью моделей знаний и их использование для построения обобщающей теории объекта, его преобразования или управления им.

Моделирование — циклический процесс. Это означает, что за первым четырехэтапным циклом может последовать второй, третий и т. д. При этом знания об исследуемом объекте расширяются и уточняются, а исходная модель постепенно совершенствуется. Недостатки, обнаруженные после первого цикла моделирования, обусловленные малым знанием объекта или ошибками в построении модели, можно исправить в последующих циклах.

Сейчас трудно указать область человеческой деятельности, где не применялось бы моделирование. Разработаны, например, модели производства автомобилей, выращивания пшеницы, функционирования отдельных органов человека, жизнедеятельности Азовского моря, последствий атомной войны. В перспективе для каждой системы могут быть созданы свои модели, перед реализацией каждого технического или организационного проекта должно проводиться моделирование.

3. Математические модели: искусство и наука

Математическая модель — это математическое представление реальности.

Математическое моделирование — процесс построения и изучения математических моделей.

Все естественные и общественные науки, использующие математический аппарат, по сути, занимаются математическим моделированием: заменяют реальный объект его математической моделью и затем изучают последнюю.

Математическое моделирование представляет собой сложный симбиоз науки и искусства, предполагающий, с одной стороны, энциклопедическое знание почти всех разделов современной математики, а с другой - тонкое интуитивное восприятие исследуемых процессов и систем.

Теоретические основы математического моделирования находятся в стадии столь высокого абстрагирования и гипертрофированной общности, что почти не допускают извлечения какого-либо конструктивного результата, полезного для практики. Иными словами, *"легко из дома реальности зайти в лес математики, но мало кто умеет вернуться назад"* [Штейнгаус].

В связи с этим большая часть авторов, старающихся приобщить новые поколения ученых к науке моделирования, предпочитает иной, вполне проторенный путь написания книг. Этот путь представляет собой формирование эклектической смеси неглубоких срезов из различных разделов прикладной математики - дифференциальных уравнений, теории случайных процессов, алгебраических, логических или топологических методов и т.п. Такого рода литература может быть полезна для развития математической эрудиции и общей ориентации в практике моделирования.

И, наконец, еще один вид литературы, относящейся к математическому моделированию, связан с узко специализированными монографиями, ориентированными на глубокое и конструктивное изучение конкретных типов моделей.

Классификация математических моделей. Априорный выбор математической модели предполагает наличие у разработчика модели, с одной стороны, глубоких знаний об изучаемом процессе или системе, а с другой - хотя бы общих представлений о существующих в настоящее время математических моделях. Упорядочение таких представлений осуществляется на основе *классификаций*. В зависимости от выбранного классификационного показателя математичсекие модели разделяются на:

статические и динамические,
линейные и нелинейные,
сосредоточенные и распределенные,
дискретные и непрерывные,
детерминированные и стохастические и т.п.

Иногда простейшая классификация оказывается не достаточно полной. Так, например, множество моделей, отражающих неопределенность ситуации, можно разделить на:

- стохастические,
- нечеткие
- хаотические.

Классификация моделей в основном повторяет классификацию систем. Однако это вовсе не означает, что тип модели должен соответствовать типу моделируемой системы. Так, например, движение любой материальной точки представляет собой непрерывный во времени процесс; в то же время в качестве его модели на практике чаще всего используется дискретная во времени модель, позволяющая выполнять расчеты с учетом специфики используемой цифровой техники.

Наряду с формальной классификацией, модели различаются по способу представления объекта на *структурные или функциональные* модели

Структурные модели представляют объект как систему со своим устройством и механизмом функционирования. *Функциональные модели* не используют таких представлений и отражают только внешне воспринимаемое поведение (функционирование) объекта. В их предельном выражении они называются также моделями «чёрного ящика». Возможны также комбинированные типы моделей, которые иногда называют моделями «*серого ящика*».

Например, бихевиоризм изучал психологию человека исключительно на основе функциональных моделей.

Пример 1. Рассмотрим механическую систему, состоящую из пружины, закрепленной с одного конца, и груза массой *m*, прикрепленного к свободному концу пружины. Будем считать, что груз может двигаться только в направлении оси пружины (например, движение происходит вдоль стержня). Построим математическую модель этой системы.

Будем описывать состояние системы расстоянием ***x*** от центра груза до его положения равновесия. Опишем взаимодействие пружины и груза с помощью *закона Гука* $F = -kx$ после чего воспользуемся вторым законом Ньютона, чтобы выразить его в форме дифференциального уравнения:

$$m\frac{d^2x}{dt^2} = -kx$$

Полученное уравнение описывает математическую модель рассмотренной физической системы. Эта модель называется «гармоническим осциллятором».

По формальной классификации эта модель линейная, детерминисткая, динамическая, сосредоточенная, непрерывная. В процессе её построения мы сделали множество допущений (об отсутствии внешних сил, отсутствии трения, малости отклонений и т. д.), которые в реальности не выполняются.

По отношению к реальности это, чаще всего, модель типа *упрощение* («опустим для ясности некоторые детали»), поскольку опущены некоторые существенные универсальные особенности (например, диссипация). В некотором приближении (скажем, пока отклонение груза от равновесия невелико, при малом трении, в течение не слишком большого времени и при соблюдении некоторых других условий), такая модель достаточно хорошо описывает реальную механическую систему, поскольку отброшенные факторы оказывают пренебрежимо малое влияние на её поведение. Однако модель можно уточнить, приняв во внимание какие-то из этих факторов. Это приведет к новой модели, с более широкой (хотя и снова ограниченной) областью применимости.

Жёсткие и мягкие модели. Гармонический осциллятор — пример так называемой «жёсткой» модели. Она получена в результате сильной идеализации реальной физической системы. Для решения вопроса о её применимости необходимо понять, насколько существенными являются факторы, которыми мы пренебрегли. Иными словами, нужно исследовать *«мягкую» модель*, получающуюся малым возмущением «жёсткой». Она может задаваться, например, следующим уравнением:

$$m\frac{d^2x}{dt^2} = -k \cdot x + \varepsilon \cdot f(x, \frac{dx}{dt}).$$

Здесь *f* — некоторая функция, в которой может учитываться сила трения или зависимость коэффициента жёсткости пружины от степени её растяжения, ε — некоторый малый параметр. Явный вид функции *f* нас в данный момент не интересует. Если мы докажем, что поведение мягкой модели принципиально не отличается от поведения жёсткой (вне зависимости от явного вида возмущающих факторов, если они достаточно малы), задача сведется к исследованию жёсткой модели. В противном случае применение результатов, полученных при изучении жёсткой модели, потребует дополнительных исследований. Например, решением уравнения гармонического осциллятора являются функции вида $x(t) = A\sin\sqrt{kt} + B\cos\sqrt{kt}$, то есть колебания с постоянной амплитудой.

Следует ли из этого, что реальный осциллятор будет бесконечно долго колебаться с постоянной амплитудой? Нет, поскольку рассматривая систему со сколь угодно малым трением (всегда присутствующим в реальной системе), мы получим затухающие колебания. Поведение системы качественно изменилось.

Если система сохраняет свое качественное поведение при малом возмущении, говорят, что она *структурно устойчива*. Гармонический осциллятор — пример структурно-неустойчивой (негрубой) системы. Тем не менее, эту модель можно применять для изучения процессов на ограниченных промежутках времени.

Пример 2. Модель Мальтуса. Скорость роста пропорциональна текущему размеру *популяции*. Она описывается дифференциальным уравнением

$\dot{x} = \alpha x$,

где α — некоторый параметр, определяемый разностью между рождаемостью и смертностью. Решением этого уравнения является экспоненциальная функция $x(t) = x_0 e^{\alpha t}$.

Если рождаемость превосходит смертность ($\alpha > 0$), размер популяции неограниченно и очень быстро возрастает.

Понятно, что в действительности этого не может происходить из-за ограниченности ресурсов. При достижении некоторого критического объёма популяции модель перестает быть адекватной, поскольку не учитывает ограниченность ресурсов.

Томас Роберт Мальтус (*Thomas Robert Malthus*, своё среднее имя он обычно опускал; 1766—1834) — английский священник и учёный, демограф и экономист, автор теории, согласно которой неконтролируемый рост народонаселения должен привести к голоду на Земле.

Уточнением модели Мальтуса может служить логистическая модель, которая описывается дифференциальным уравнением Ферхюльста

$$\dot{x} = \alpha(1 - \frac{x}{x_s})x,$$

где x_s — «равновесный» размер популяции, при котором рождаемость в точности компенсируется смертностью. Размер популяции в такой модели стремится к равновесному значению x_s, причем такое поведение структурно устойчиво.

Пример 3. Система хищник-жертва. Допустим, что на некоторой территории обитают два вида животных: кролики (питающиеся растениями) и лисы (питающиеся кроликами). Пусть число кроликов x, число лис y. Используя модель Мальтуса с необходимыми поправками, учитывающими поедание кроликов лисами, приходим к следующей системе, носящей имя *модели Лотки — Вольтерра:*

$$\begin{cases} \dot{x} = (\alpha - cy)x; \\ \dot{y} = (-\beta + dx)y. \end{cases}$$

Эта система имеет равновесное состояние, когда число кроликов и лис постоянно. Отклонение от этого состояния приводит к колебаниям численности кроликов и лис, аналогичным колебаниям гармонического осциллятора. Как и в случае гармонического осциллятора, это поведение не является структурно устойчивым: малое изменение модели (например, учитывающее ограниченность ресурсов, необходимых кроликам) может привести к качественному изменению поведения. Например, равновесное состояние может стать устойчивым, и колебания численности будут затухать. Возможна и противоположная ситуация, когда любое малое отклонение от положения равновесия приведет к катастрофическим последствиям, вплоть до полного вымирания одного из видов. На вопрос о том, какой из этих сценариев реализуется, модель Вольтерра — Лотки ответа не дает: здесь требуются дополнительные исследования.

Разработка модели. Разработка математической модели, как правило, включает в себя три основных этапа

- выбор структуры,
- определение параметров
- проверку подобия.

В общем случае моделирование осуществляется итерационно. Это связано с тем, что практически невозможно оценить качество выбранной непараметрической структуры модели без оценки ее параметров.

Осуществив априорный выбор структуры, разработчик модели переходит к определению ее параметров. В случае, когда имеется набор численных данных, полученных в результате наблюдения за функционированием реальной системы-прототипа или ее экспериментального аналога, искомые параметры находятся путем применения алгоритмов обработки указанной информации. В частности, широко используется математический аппарат статистического оценивания.

Последний этап создания модели связан с проверкой ее адекватности, или подобия моделируемой системе. При этом адекватность понимается в крайне узком смысле этого слова, а именно в контексте близости изучаемых свойств системы и аналогичных характеристик модели. Например, при исследовании выходных характеристик системы формируется мера близости между конкретными выходными показателями реальной системы и ее модели.

Очевидно, что данная мера представляет собой реализацию некоторой случайной величины. Определив функцию распределения этой величины, можно найти по соответствующим таблицам математической статистики критические значения выбранной меры, отвечающие a'priori выбранному уровню доверия разработчика модели своему детищу. Если уровень доверия оказывается ниже критического, необходимо перейти либо к первому этапу и скорректировать структуру модели, либо ко второму этапу сбора и обработки экспериментальных данных. И все начинается сначала...

NB!: "*Цифры не управляют миром, но они показывают, как управляется мир". И. В. Гете*

Вопросы для самопроверки:

1. Что называется моделью?
2. Перечислите основные виды моделирования.
3. Что такое ментальное моделирование?
4. Назовите виды моделей по способу отображения реальности.
5. Что представляют собой эвристические модели?
6. Чем отличаются натурные модели от реальных объектов?
7. Что такое математическое моделирование?
8. Назовите основные этапы моделирования.
9. Приведите классификацию математических моделей.
10. Что такое структурны и функциональные модели?

ЛИТЕРАТУРА:

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных / Под ред. С.А.Айвазяна.- М.: Финансы и статистика, 1993.- 471с.

2. Директор С., Рорер Р. Введение в теорию систем/Пер. с англ. под ред. Н.П. Бусленко. - М.: Мир, 1974. - 464с.

3. Калинин В.Н., Резников Б.А., Варакин Е.И. Теория систем и оптимального управления. - Л.: ВИКИ им. А.Ф. Можайского, 1979. - 319с.

4. Бусленко Н.П. Моделирование сложных систем. - М.: Наука, 1978. - 399с.

5. Егоренков Д.Л., Фрадков А.Л., Харламов В.Ю. Основы математического моделирования с примерами на языке MATLAB. - СПб БГТУ, 1996. - 192с.

6. Ли Р. Оптимальные оценки, определение характеристик и управление/Пер. с англ. под ред. Я.З. Цыпкина. - М.: Наука, 1966.-176с.

7. Математическое моделирование/Под ред. Д.Эндрюса, Р.Маклоуна. Пер. с англ. - М.: Мир, 1979. - 277с.

8. Месарович Д., Такахара Я. Общая теория систем: математические основы/Пер. с англ. под ред. С.В. Емельянова. - М.: Мир, 1978. - 312с.

9. Моисеев Н.Н. Математические задачи системного анализа. - М.: Наука, 1981. - 488с.

10. Морозов Л.М., Петухов Г.Б., Сидоров В.Н. Методологические основы теории эффективности: Учебное пособие. - Л.: ВИКИ им. А.Ф.Можайского, 1982. - 236с.

11. Николис Г., Пригожин И. Самоорганизация в неравновесных структурах/Пер. с англ. - М.: Мир, 1979. - 327с.

12. Перегудов Ф.И., Тарасенко Ф.П. Введение в системный анализ. - М.: Высшая школа, 1989. - 367с.

13. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. - М.: Наука, 1979. - 244с.

14. Цурков В.Н. Агрегирование данных при решении динамических задач большой размерности. - М.: Наука, 1987. - 484с.

15. Штейнгауз Г. Математический калейдоскоп/Пер. с польск. - М.: ГИТТЛ, 1949. - 143с.

16. *Самарский А.А., Михайлов А.П.* Математическое моделирование: Идеи. Методы. Примеры. — М: Наука, 1997. — 320 с.

ЛЕКЦИЯ 4

СИСТЕМНЫЙ АНАЛИЗ ДАННЫХ

1. Введение. Системный подход: взаимосвязь всего со всем

NB!

Определение. *Системный анализ* — научный метод познания, представляющий собой последовательность действий по установлению структурных связей между переменными или постоянными элементами исследуемой системы. Опирается на комплекс общенаучных, экспериментальных, естественнонаучных, статистических, математических методов.

Системный анализ (СА) является компонентом Computer Science, он возник в эпоху разработки компьютерной техники и напрямую связан с уровнем развития информационных технологий.

Н. Н. Моисеев приводит, по его выражению, довольно узкое определение системного анализа: «Системный анализ — это совокупность методов, основанных на использовании ЭВМ и ориентированных на исследование сложных систем — технических, экономических, экологических и т.д.» [Моисеев].

Результатом системных исследований является, как правило, выбор вполне определенной альтернативы: плана развития региона, параметров конструкции и т. д. Поэтому истоки системного анализа, его методические концепции лежат в тех дисциплинах, которые занимаются проблемами принятия решений: *исследование операций* и *общая теория управления.*

Ценность системного подхода состоит в том, что рассмотрение категорий системного анализа создает основу для логического и последовательного подхода к проблеме принятия решений. Эффективность решения проблем с помощью системного анализа определяется структурой решаемых проблем.

В зависимости от качества структуризации данных, проблемы СА можно разделить на три класса:

- хорошо структурированные (*well-structured*), или количественно сформулированные проблемы, в которых существенные зависимости выяснены очень хорошо;
- слабо структурированные (*ill-structured*), или смешанные проблемы, которые содержат как качественные элементы, так и малоизвестные, неопределенные стороны, которые имеют тенденцию доминировать;
- неструктурированные (*unstructured*), или качественно выраженные проблемы, содержащие лишь описание важнейших ресурсов, признаков и характеристик, количественные зависимости между которыми совершенно неизвестны.

Системность – это важнейший атрибут научного мышления, проявляющаяся в способности видеть взаимосвязанное единство в совокупности разрозненных фактов, событий, элементов.

Примечание. *В соответствии с законами "мэрфологии" основные постулаты развитой теории систем имеют следующий вид:*

1. Все - система.

2. Все - часть еще большей системы.

3. Вселенная бесконечно систематизирована как снизу вверх (все более крупные системы), так и сверху вниз (меньшие системы).

4. Все системы бесконечно сложны. (Иллюзия простоты возникает из-за сосредоточения внимания на одной или нескольких переменных.)

NB! Основным постулатом системного анализа является необходимость исследования не только тех или иных явлений, но и совокупности всевозможных сопутствующих им взаимосвязей, без которых полученные решения не будут ни достаточно полными, ни достаточно корректными с точки зрения понимания изучаемого вопроса. Именно взаимосвязанность составных частей является основной характеристикой одного из важнейших объектов научных исследований - системы.

Понятие *"система"* представляет собой базовую научную категорию в самых разнообразных областях знаний. Существует множество определений этого понятия, различающихся между собой той или иной степенью неудачности.

Само слово "система" произошло от греческого σισθεμασ - целое, составленное из частей.

Наиболее общее, философское определение утверждает:

система — это множество элементов, находящихся в отношениях и связях друг с другом, которое образует определенную целостность, единство.

Характерными свойствами системы являются: **NB!**

- *целостность,*
- *структурированность,*
- *целенаправленность (*для большинства искусственных систем).

Свойство *целостности* говорит о том, что система может быть выделена из окружающей ее внешней среды, хотя и находится во взаимодействии с нею.

Структурированность означает, что система может быть разделена на подсистемы или составные элементы, связанные и взаимодействующие друг с другом.

Свойство *целенаправленности* предполагает задание некоторой цели, достижение которой определяет предназначение системы.

Формализация понятия "система" приводит к более конкретным, хотя и более узким определениям.

В частности, в кибернетике система задается в виде математических моделей $S = \{U, Y\}$ или $S = \{U, X, Y\}$, где $\mathbf{U}$ - множество входов, $\mathbf{X}$ - множество состояний, $\mathbf{Y}$ - множество выходов. **NB!**

Норберт Винер (26 ноября 1894, Колумбия, штат Миссури, США — 18 марта 1964, Стокгольм, Швеция) — американский учёный, выдающийся математик и философ, основоположник кибернетики и теории искусственного интеллекта.

Таким образом, систему можно трактовать, как некоторый *"черный ящик"*, преобразующий входные сигналы (данные) в выходные в интересах некоторой априори заданной цели.

В случае, когда функционирование системы изучается с точки зрения изменений ее состояния во времени, $S:\{X(t),\ t \in T\}$, где T - интервал времени, система называется динамической.

Если каждой входной функции $u(t)$ соответствует единственная выходная функция $\mathbf{y(t)}$, то такая система называется *функциональной.*

Иными словами, если $u_1(t) = u_2(t)$ при $t \le \tau$, то соответствующие им выходные сигналы будут удовлетворять требованию $\mathbf{y_1(\tau) = y_2(\tau)}$. В противном случае ее относят к категории *неопределенных*.

Важнейшим свойством, присущим реальным системам, является *причинность их эволюции*. Это означает, что настоящее состояние системы определяется ее прошлыми состояниями и не зависит от будущего.

Представленная формализация понятия системы носит явно выраженный функциональный характер и не затрагивает ее структуру. В то же время в задачах синтеза важнейшую роль играет именно структура системы, которую и требуется определить.

В связи с этим обобщенная модель системы еще более усложняется и приобретает вид

NB!

$$S = \{A, R, U, X, Y, G\}, \qquad (1)$$

где *A* - множество элементов системы; $\mathbf{R}$ - матрица отношений между элементами системы, заданными на *A*; $G = G(A, X)$ - матрица отношений между множествами $\mathbf{A}$ и $\mathbf{X}$ (отношения *эмерджентности*). Составляющая часть системы $\{A, R\}$ называется ее структурой, а $\{U, X, Y,\}$ - ее программой функционирования. Таким образом, система представляет собой объединение этих двух составляющих, а ее единство определяется отношением эмерджентности *G*.

Заметим, что понятие элемента системы является первичным и как всякое первичное представление очень сложно формализуется, но достаточно очевидно на уровне "здравого смысла": элемент — это составляющая часть системы на последней ступени ее детализации. Декомпозицию можно проводить до атомарного уровня и ниже; здесь, как и вообще в жизни, следует вовремя остановиться, не утратив функциональной осмысленности.

Следует заметить, что категория "система", несмотря на кажущуюся простоту, достаточно загадочна и малообъяснима с позиций материализма. Так, например, расширение системы вполне может привести к диалектическому скачку, описанному Георгом Вильгельмом Фридрихом Гегелем в XIX веке, в результате которого образуется новое качество. При этом новое качество может быть совсем не присущим ни одному элементу, образующему эту систему.

Примером этого таинства может служить, например, книга - последовательность знаков (букв, пробелов, символов) в совокупности составляет носитель новой (как надеются ее авторы) информации.

Еще один удивительный пример - *синергетические системы* [Николис, Хакен]. Здесь вопреки "великому и ужасному» второму закону термодинамики моделировались и наблюдались спонтанные процессы самоорганизации!

Возвращаясь к обсуждению свойств систем, отметим, что важнейшее из них (с точки зрения практического человека) - целенаправленность - наиболее характерно для *homo sapiens* и связанных с его деятельностью искусственных структур.

В отношении природных систем сказать что-либо однозначное об их целенаправленности достаточно сложно - есть ли цель у неба? у моря? у погоды (кроме как испортить в Санкт-Петербурге выходной день)? Сам факт наличия закономерностей в природе (законы той же диалектики, законы физики и т.п.) указывает на то, что в принципе какая-то цель (или какие-то цели) вполне может существовать. Эту цель материалисты называют развитием, совмещая результат с процессом его достижения, идеалисты - Божественным промыслом, самопознанием Идеи и т.п. Однако назвать что-либо еще не означает это понять.

В искусственных системах все гораздо понятнее: цель - это *результат*. Если получен результат, то цель достигнута. Достижение цели, в свою очередь, состоит из решения поставленной задачи (или совокупности задач, образующих проблему). Результат не возникает из ничего, для его достижения всегда необходимо наличие определенных *ресурсов*.

Преобразование ресурса в результат, обеспечивающий достижение цели, осуществляется упорядоченной совокупностью взаимосвязанных действий, образующих *операцию*.

NB! Наличие последнего понятия позволило получить еще одно, вполне прагматическое определение системы [Морозов]: система - это множество взаимосвязанных материальных объектов, непосредственно участвующих в процессе выполнения операции.

Взаимосвязи приведенных понятий можно представить в виде структурной схемы [10], показанной на рис. 1.

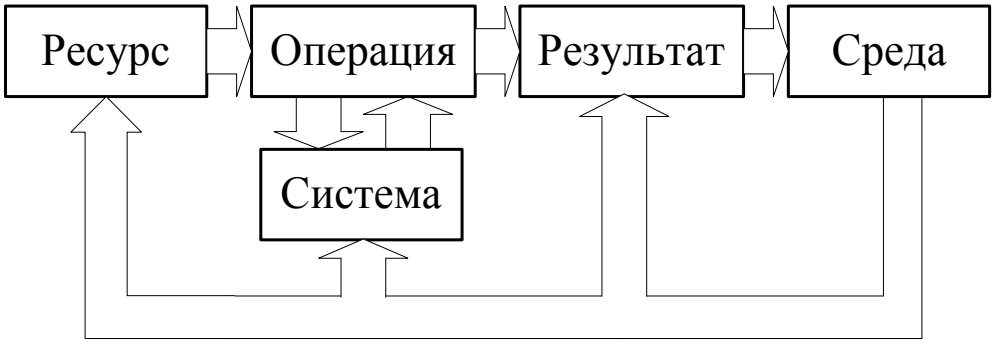


Рис. 1. Взаимосвязь основных понятий системного анализа

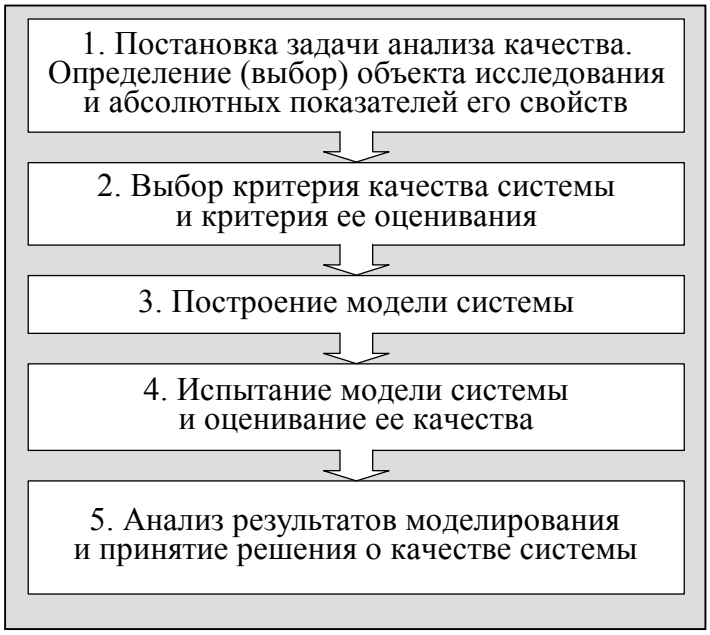


Рис. 2. Прямая задача: анализ качества системы

В большинстве исследованиях, как правило, приходится иметь дело с так называемыми *"сложными"*, или *"большими"* системами, рассмотрению которых посвящен следующий раздел лекции. В таких системах особенно подчеркнуто их важнейшее свойство - взаимосвязанность элементов.

NB! В дальнейшем будем различать *две основные задачи теории систем: прямую и обратную* [9]. Структурные схемы решений указанных задач приведены на рис. 2 и 3. Прямая задача, или *задача анализа*, состоит в изучении *результатов функционирования системы и определении условий ее применения*.

Обратная задача, или *задача синтеза*, связана с определением структуры, параметров и свойств системы в заданном диапазоне возможных условий. При этом предполагается, что в результате синтеза в конечном итоге будет получена система, отвечающая выбранным критериям качества.

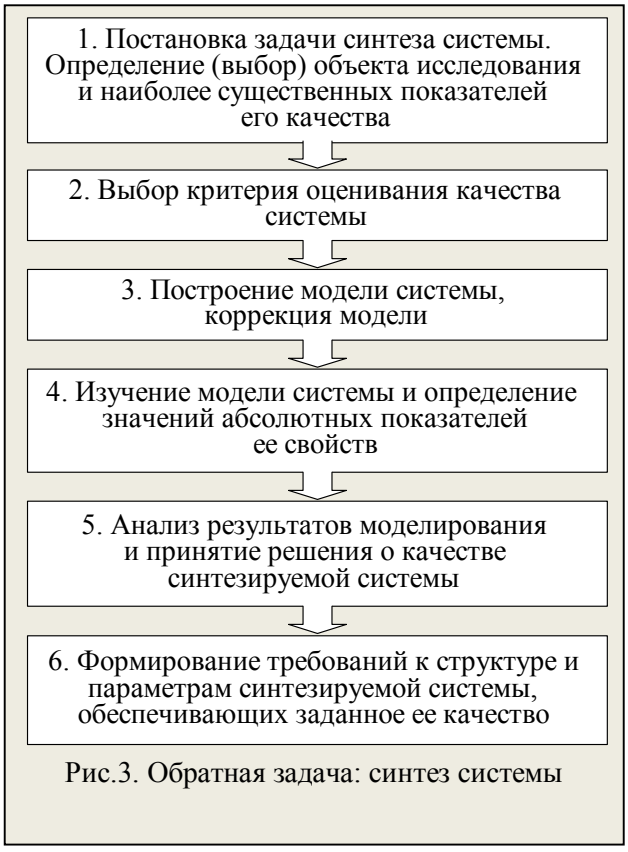


Рис.3. Обратная задача: синтез системы

Основополагающим этапом понимания любого предмета является *классификация*. Однако задача классификации систем достаточно сложна в силу крайне высокого уровня общности самого этого понятия. Тем не менее, эту работу надо делать: проведение классификации позволяет исследователю сузить область анализа и подобрать наиболее подходящую для данной задачи математическую модель.

В качестве первичных классификационных признаков будем использовать, следуя [10], элементы модели (1). В частности, если осуществлять классификацию множества систем $\{S\}$ на основе природы элементов $a \in A$, то все системы можно разделить на физические, технические, социальные, военные и т.п.

Классификация по свойствам отношений $\mathbf{r} \in \mathbf{R}$ приводит к разделению множества $\{S\}$ на системы с постоянной $(\mathbf{R} = \mathbf{const})$ и переменной $(\mathbf{R} = \mathbf{var}, \mathbf{R} = \mathbf{R(t)})$ структурой.

Классификация, основанная на свойствах входа-выхода системы, позволяет разделить множество $\{S\}$ по взаимодействию со средой

на открытые $(\mathbf{U} \neq \varnothing \cap \mathbf{Y} \neq \varnothing)$ и закрытые системы $(\mathbf{U} = \varnothing \cap \mathbf{Y} = \varnothing)$,

на активные $(\mathbf{Y} \neq \varnothing)$ и пассивные $(\mathbf{Y} = \varnothing)$,

а также по числу входов-выходов (например, система с *m* входами и *n* выходами).

Классификация по свойствам множества $\mathbf{X}$ разделяет множество систем в зависимости:

- от характера возможных состояний - на дискретные и непрерывные,
- от изменения состояния во времени - на статистические $(\mathbf{X} = \mathbf{const})$ и динамические $(\mathbf{X} = \mathbf{X(t)})$,
- в зависимости от степени учета случайных факторов - на детерминированные и стохастические.

И, наконец, классификация на основе отношения эмерджентности G позволяет разделить все системы на регулярные и нерегулярные в зависимости от однозначности отображения множества элементов A на множество возможных состояний X.

Заметим, что при решении прикладных задач часто используются некоторые установившиеся, общепринятые разбиения систем на составные части. Так, система управления разбивается на управляющую и управляемую подсистемы, а управляющая система - на подсистему наблюдения и собственно управления.

И в заключение этого раздела приведем диалог Конфуция с учеником, отражающий отношение великого мудреца к системному анализу.

"-Ты считаешь меня многоученым? - спросил Конфуций ученика. - А разве нет? - ответил тот. - Нет, - сказал Конфуций, - я лишь связываю все воедино".

2. Сложные системы: Проклятие размерности, декомпозиция

Классической проблемой почти всех прикладных наук является проблема *сложных* систем.

Как правило, это понятие в каждом конкретном случае уточняется в зависимости от контекста решаемой проблемы.

NB! В общем случае *под сложной системой* понимают многофункциональную систему с многомерной, многосвязной, неоднородной структурой, нестационарно изменяющейся во времени и содержащей существенные неопределенности в описании. Иногда, "для букета", добавляют некорректность задания, недифференцируемость протекающих в ней процессов, ограниченную наблюдаемость, слабую управляемость и т.п.

Можно предложить альтернативное определение: *сложная система — это система, не допускающая красивого математического описания*.

Вывод из представленного выше определения достаточно очевиден - необходимо по возможности упростить и модель исследуемой системы, и саму решаемую задачу. Соответствующие математические технологии описаны, напри-

мер, в [Егор, Месар, Моисеев, Перегуд]. Проблема состоит в том, чтобы сделать это корректно!

NB! Введение корректных ограничений и упрощений, строгое применение техники декомпозиции - задачи, требующие особого внимания и предельной аккуратности. Именно здесь возникают серьезные вопросы, а иногда и проблемы при защите полученных результатов.

Идея декомпозиции достаточно очевидна - разделить изучаемую систему на подсистемы, а решаемую задачу - на подзадачи, каждая из которых допускает более или менее самостоятельное исследование. Если вычлененная подсистема слишком сложна, то продолжают процесс разбиения до тех пор, пока не получат подсистему, допускающую определенное решение.

По существу, реализуется иерархическая, или многоуровневая структуризация исходной задачи (системы). Для формализованного решения задачи декомпозиции удобно использовать математический аппарат теории графов и технологию структурного программирования.

NB! Как правило, сложная система допускает *несколько вариантов декомпозиции*, что связано с наличием различных подходов к задаче анализа ее функционирования. В результате возникают неоднозначность, субъективизм выбора, многокритериальность и другие неприятности, существенно затрудняющие получение строгого решения. При этом не последнюю роль приобретают интуиция и опыт исследователя.

По данным психологов, человек может мысленно охватить структуру декомпозированной системы, если на каждом уровне возникает не более 5±2 подзадач [ЕгорФрадк].

Научная методология постановки и решения задач исследования сложных систем получила наименование ***"системный анализ"***. По своей природе системный анализ представляет собой некоторое обобщение различных методических приемов, возникающих при решении конкретных естественно-научных, социальных, военных, технических и других задач. Несмотря на содержательную разнородность таких задач, системный анализ позволяет существенно унифицировать технологию их решения. В частности, в работе [Егор] представлен вариант общей схемы алгоритма решения задачи исследования системы (рис. 1).

Следует заметить, что одной из типичных проблем, возникающих при исследовании сложных систем, является так называемое "проклятие размерности" [Ли]. В соответствии с законами Мэрфи, *"задача, имеющая размерность меньше трех, тривиальна, а имеющая размерность больше восьми - не имеет решения".*

Очевидной неприятностью, вытекающей из высокой размерности решаемой задачи, является нелинейный рост требований к вычислительным ресурсам используемой ЭВМ. Так, например, объем вычислений, необходимых для реализации алгоритма следящего наблюдателя, растет пропорционально третьей степени от размерности вектора состояния контролируемого объекта.

К сожалению, возникающие в результате роста размерности задачи проблемы не исчерпываются чисто техническими аспектами; в задачах с высокой размерностью возникают многочисленные "таинственные" явления, объяснение

которых представляет собой самостоятельную, отнюдь не тривиальную задачу. Как замечено в мэрфологическом принципе Шательера, *сложные системы имеют тенденцию противопоставлять себя своим же функциям.*

В качестве примера приведем задачи теории наблюдения динамических систем, в которых рост числа контролируемых параметров увеличивает вероятность появления таких переменных, которые слабо зависят от изменений входных переменных (измерений).

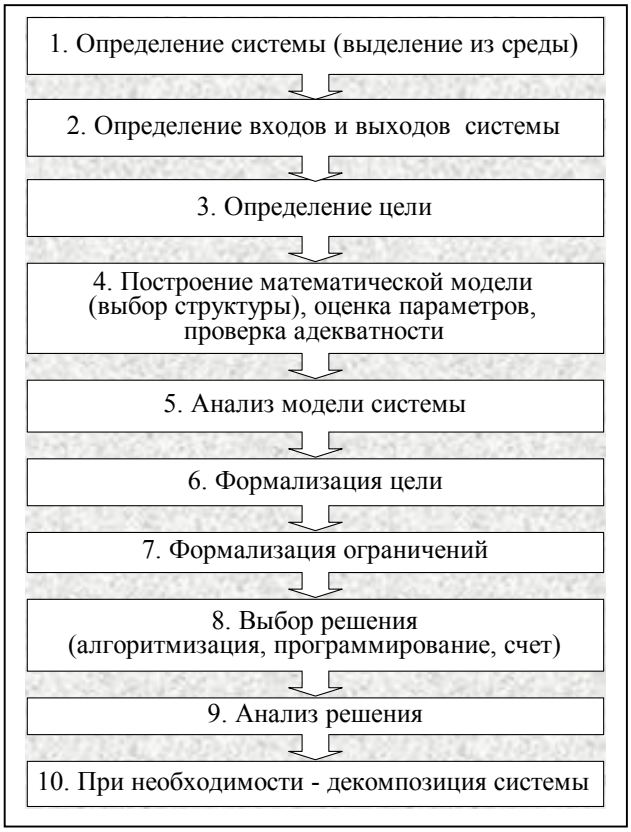


Рис. 4. Общая схема алгоритма решения задачи исследования системы

В результате соответствующая градиентная матрица оказывается вырожденной или плохо обусловленной, что приводит к полной или частичной потере наблюдаемости.

Аналогичные проблемы возникают в задачах управления большеразмерными системами, в задачах прогнозирования, идентификации и т.п.

Мэрфологический принцип неопределенности утверждает, что *системы имеют тенденцию расти и, по мере роста, взаимораствориться*. Другие формулировки звучат не менее пессимистично: *"Сложные системы приводят к неожиданным последствиям"* или *"Совокупное поведение больших систем предсказать нельзя"*.

Для борьбы с "проклятьем размерности" используется разнообразный математический аппарат, охватывающий широкий спектр вычислительных технологий от методов агрегирования сложных систем (задача, обратная декомпозиции) [Бусленко, Цурков] до алгоритмов тихоновской или иной регуляризации [Тихон]. Выбор наиболее подходящей технологии в каждом конкретном случае индивидуален и лежит в области системотехнического искусства, приобретаемого долгим и нелегким опытом.

Остается надеяться, что если не рядовой разработчик, то хотя бы руководитель проекта таким опытом обладает.

3. Методы решения задач системного анализа

Для решения хорошо структурированных количественно выражаемых проблем используется известная методология *исследования операций*, которая состоит в построении адекватной математической модели (например, задачи линейного, нелинейного, динамического программирования, задачи теории массового обслуживания, теории игр и др.) и применении методов для отыскания оптимальной стратегии управления целенаправленными действиями.

Системный анализ предоставляет к использованию в различных науках, системах следующие системные методы и процедуры:

- абстрагирование и конкретизация
- анализ и синтез, индукция и дедукция

- формализация и конкретизация
- композиция и декомпозиция
- линеаризация и выделение нелинейных составляющих
- структурирование и реструктурирование
- макетирование
- реинжиниринг
- алгоритмизация
- моделирование и эксперимент
- программное управление и регулирование
- распознавание и идентификация
- кластеризация и классификация
- экспертное оценивание и тестирование
- верификация

и другие методы и процедуры.

Вопросы для самопроверки:

1. Приведите различные определения системного анализа.
2. Назовите цель системного анализа.
3. Дайте классификацию СА в зависимости от уровня структуризации данных.
4. Приведите основной постулат СА.
5. Назовите основные свойства системы.
6. Приведите формализованное описание системы;
7. Какая система называется функциональной? неопределенной?
8. Что называется отношением *эмерджентности*?
9. Опишите взаимосвязь основных понятий системного анализа.
10. Опишите две основные задачи теории систем: прямую и обратную.

ЛИТЕРАТУРА:

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных / Под ред. С.А. Айвазяна.- М.: Финансы и статистика, 1993.- 471с.

2. Директор С., Рорер Р. Введение в теорию систем/Пер. с англ. под ред. Н.П. Бусленко. - М.: Мир, 1974. - 464с.

3. Калинин В.Н., Резников Б.А., Варакин Е.И. Теория систем и оптимального управления. - Л.: ВИКИ им. А.Ф. Можайского, 1979. - 319с.

4. Бусленко Н.П. Моделирование сложных систем. - М.: Наука, 1978. - 399с.

5. Егоренков Д.Л., Фрадков А.Л., Харламов В.Ю. Основы математического моделирования с примерами на языке MATLAB. - СПб БГТУ, 1996. - 192с.

6. Ли Р. Оптимальные оценки, определение характеристик и управление/Пер. с англ. под ред. Я.З. Цыпкина. - М.: Наука, 1966.-176с.

7. Математическое моделирование/Под ред. Д.Эндрюса, Р.Маклоуна. Пер. с англ. - М.: Мир, 1979. - 277с.

8. Месарович Д., Такахара Я. Общая теория систем: математические основы/Пер. с англ. под ред. С.В. Емельянова. - М.: Мир, 1978. - 312с.

9. Моисеев Н.Н. Математические задачи системного анализа. - М.: Наука, 1981. - 488с.

10. Морозов Л.М., Петухов Г.Б., Сидоров В.Н. Методологические основы теории эффективности: Учебное пособие. - Л.: ВИКИ им. А.Ф.Можайского, 1982. - 236с.

11. Николис Г., Пригожин И. Самоорганизация в неравновесных структурах / Пер. с англ. - М.: Мир, 1979. - 327с.

12. Перегудов Ф.И., Тарасенко Ф.П. Введение в системный анализ. - М.: Высшая школа, 1989. - 367с.
13. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. - М.: Наука, 1979. - 244с.
14. Цурков В.Н. Агрегирование данных при решении динамических задач большой размерности. - М.: Наука, 1987. - 484с.
15. Штейнгауз Г. Математический калейдоскоп/Пер. с польск. - М.: ГИТТЛ, 1949. – 143с.

ЛЕКЦИЯ 5

ЭЛЕМЕНТЫ ТЕОРИИ ДИНАМИЧЕСКИХ СИСТЕМ

1. Динамическая система и ее математическая модель

Определение. Под *ДС* понимают любой объект или процесс, для которого однозначно определено понятие состояния как совокупности некоторых величин в данный момент времени и задан закон, который описывает изменение (эволюцию) начального состояния с течением времени. **NB!**

Закон эволюции позволяет по начальному состоянию прогнозировать будущее состояние ДС системы.

ДС – это механические, физические, химические и биологические объекты, состояние которых изменяется во времени в соответствии с конкретными алгоритмами. Описания ДС для задания закона эволюции также разнообразны: с помощью дифференциальных уравнений, дискретных отображений, теории графов, теории марковских цепей и т.д. Выбор одного из способов описания задает конкретный вид математической модели соответствующей ДС.

Математическая модель ДС считается заданной, если введены параметры (координаты) системы, определяющие однозначно ее состояние, и указан закон эволюции. В зависимости от степени приближения одной и той же системе могут быть поставлены в соответствие различные математические модели.

Исследование реальных систем сводится к изучению математических моделей, совершенствование и развитие которых определяются анализом экспериментальных и теоретических результатов при их сопоставлении. В связи с этим изучение ДС всегда ассоциировано с исследованием ее математической модели.

Исследуя одну и ту же ДС (к примеру, движение маятника), в зависимости от степени учета различных факторов мы получим различные математические модели. В качестве примера рассмотрим модель нелинейного консервативного осциллятора:

$$\frac{d^2x}{dt^2} + \sin x = \ddot{x} + \sin x = 0$$

Функция sin x аналитическая, и ее можно разложить в ряд Тейлора:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ... =$$

$$= \sum_{n=1}^{\wp} \frac{x^{4n+1}}{(4n+1)!} - (\sum_{n=1}^{\wp} \frac{x^{4n-1}}{(4n-1)!})$$

При малых $x << 1$ $\sin x \cong x$. В этом случае получаем самую простую модель математического маятника $\ddot{x} + x = 0$.

С увеличением x требуется учет второго, третьего и т.д. членов ряда, чтобы с заданной точностью аппроксимировать sin x. Следующим приближением будет модель нелинейного маятника:

$$\ddot{x} + x - \frac{x^3}{6} = 0$$

и т.д. Для каждого конкретного значения n будем получать новую ДС, в заданном приближении описывающую процесс колебаний физического маятника

2. Кинематическая интерпретация

Рассмотрим ДС, моделируемые конечным числом обыкновенных дифференциальных уравнений. В рассматриваемом случае для определения ДС необходимо указать объект, допускающий описание состояния заданием величин x_1 , x_2 , …, x_N в некоторый момент времени $t = t_0$.

Величины x_i могут принимать произвольные значения, причем двум различным наборам величин x_i отвечают два разных состояния. Закон эволюции ДС во времени записывается системой обыкновенных дифференциальных уравнений

$$\frac{dx_i}{dt} = \dot{x}_i = f_i(x_1,...,x_n), \quad i = 1,2,...,N \tag{1}$$

Если рассматривать величины x_1 , x_2 , …, x_N как координаты точки x в N-мерном пространстве, то получается наглядное геометрическое представление состояния ДС в виде этой точки, которую называют изображающей или *фазовой точкой. Множество фазовых точек с введенной над ним метрикой называется пространством состояний* или *фазовым пространством* ДС. Изменению состояния системы во времени отвечает движение фазовой точки вдоль некоторой линии, называемой *фазовой траекторией*. В фазовом пространстве системы уравнениями (1) определяется векторное поле скоростей, сопоставляющее каждой точке x выходящий из нее вектор скорости F(x), компоненты которого даются правыми частями уравнений (1). При этом ДС (1) может быть записана в векторной форме:

$$\dot{X} = F(X) \tag{2}$$

где F(x) – вектор-функция размерности N. Необходимо уточнить взаимосвязь понятий числа степеней свободы и размерности фазового пространства ДС.

Под *числом степеней свободы* понимается наименьшее число независимых координат, необходимых для однозначного определения состояния системы.

Под координатами первоначально понимались именно пространственные переменные, характеризующие взаимное расположение тел и объектов.

В то же время для однозначного решения соответствующих уравнений движения необходимо помимо координат задать соответствующие начальные значения *импульсов* или *скоростей*. В связи с этим система с n степенями свободы характеризуется фазовым пространством в два раза большей размерности (N = 2n).

3. Классификация динамических систем

Если ДС задана уравнением (2), то постулируется, что каждому $x(t_0)$ в фазовом пространстве ставится в соответствие состояние x(t), $t > t_0$, куда за время $t - t_0$ переместится фазовая точка, движущаяся в соответствии с уравнением (2).

В операторной форме (2) можно записать в виде

$x(t) = T_t\, x(t_0)$,

где T_t – закон (оператор) эволюции.

Если этот оператор применить к начальному состоянию $x(t_0)$, то получим x(t), то есть состояние в момент времени $t > t_0$. Так как $x(t_0)$ и x(t) принадлежат одному и тому же фазовому пространству ДС, то математики говорят в данной ситуации*: оператор T_t отображает фазовое пространство системы на себя.* В

соответствии с этим можно называть оператор T_t *оператором отображения* или просто *отображением*.

ДС можно классифицировать в зависимости от вида оператора отображения и структуры фазового пространства.

Если оператор предусматривает исключительно *линейные преобразования* начального состояния, то он называется линейным. Линейный оператор обладает свойством *суперпозиции:* T[x(t) + y(t)] = Tx(t) + Ty(t).

Если оператор нелинейный, то и соответствующая ДС называется нелинейной.

Различают непрерывные и дискретные операторы и соответственно системы с непрерывным и дискретным временем. Системы, для которых отображение x(t) с помощью оператора T может быть определено для любых $t > t_0$ (непрерывно во времени), называют также *потоками* по аналогии со стационарным течением жидкости. Если оператор отображения определен на дискретном множестве значений времени, то соответствующие ДС называют *каскадами* или *системами с дискретным временем*.

Способы задания оператора отображения T также могут различаться. Оператор T можно задать в виде дифференциального или интегрального преобразования, в виде матрицы или таблицы, в виде графика или функции и т.д.

4. Колебательные системы и их свойства

Важную группу динамических систем представляют системы, в которых возможны колебания. Колебательные системы с точки зрения их математических моделей разделяют на определенные классы.

Различают линейные и нелинейные колебательные системы, сосредоточенные и распределенные, консервативные и диссипативные, автономные и неавтономные.

Особый класс представляют так называемые автоколебательные системы. Основные свойства указанных систем подробно обсуждаются в работах по теории колебаний.

Колебательная система называется *линейной* или *нелинейной* в зависимости от того, линейна или нелинейна описывающая ее система дифференциальных уравнений. Линейные системы являются частным случаем нелинейных. Однако в силу принципиальной важности линейных систем при исследовании вопросов устойчивости колебаний, а также возможности использования принципа суперпозиции решений такая классификация оправданна.

ДС, моделируемые конечным числом обыкновенных дифференциальных уравнений, называют *сосредоточенными* или *точечными системами.* Они описываются с помощью конечномерного фазового пространства и характеризуются конечным числом степеней свободы.

Одна и та же система в различных условиях может рассматриваться либо как сосредоточенная, либо как *распределенная*.

Математические модели распределенных систем – это *дифференциальные уравнения в частных производных, интегральные уравнения или обыкновенные уравнения с запаздывающим аргументом.*

Число степеней свободы распределенной системы бесконечно, и требуется бесконечное число данных для определения ее состояния.

По энергетическому признаку ДС делятся на *консервативные* и *неконсервативные*. Консервативные системы характеризуются неизменным во времени запасом энергии. В механике их называют *гамильтоновыми.*

Для консервативных систем с n степенями свободы определяется гамильтониан системы H(p, q), где q_i – обобщенные координаты, p_i – обобщенные импульсы (mv) системы, i = 1, 2, …, n.

Гамильтониан полностью характеризует динамическую природу системы и с физической точки зрения в большинстве случаев представляет собой ее полную энергию. Эволюция во времени консервативных систем описывается уравнениями механики Гамильтона

$$\dot{q}_i = \frac{\partial H(p,q)}{\partial p_i}, \quad \dot{p}_i = \frac{\partial H(p,q)}{\partial q_i}.$$

ДС с изменяющимся во времени запасом энергии называются *неконсервативными.* Неконсервативные системы, в которых энергия уменьшается во времени из-за трения или рассеяния, называются *диссипативными.*

В соответствии с этим системы, энергия которых во времени нарастает, называются системами с отрицательным трением или отрицательной диссипацией. Такие системы можно рассматривать как диссипативные при смене направления отсчета времени на противоположное.

ДС называются *автономными*, если они не подвержены действию внешних сил, переменных во времени. Уравнения автономных систем явной зависимости от времени не содержат.

Большинство реальных колебательных систем в физике, радиофизике, биологии, химии и других областях знаний неконсервативны. Среди них выделяется особый класс *автоколебательных систем*, которые принципиально неконсервативны и нелинейны.

Автоколебательной называют ДС, преобразующую энергию источника в энергию незатухающих колебаний, причем основные характеристики колебаний (амплитуда, частота, форма колебаний и т.д.) определяются параметрами системы и в определенных пределах не зависят от выбора исходного начального состояния.

5. Фазовые портреты типичных колебательных систем

Геометрическое представление колебаний. Метод анализа колебательных процессов с помощью исследования фазовых траекторий ДС был введен в теорию колебаний Л.И. Мандельштамом и А.А. Андроновым и с тех пор стал привычным при исследовании различных колебательных явлений.

Обсудим несколько простых, но типичных примеров представления динамических процессов в виде траекторий изображающей точки в фазовом пространстве.

Консервативный осциллятор. Рассмотрим ***линейный*** осциллятор без потерь, уравнения которого можно сформулировать на примере колебательного LC-контура (рис. 1, а), предположив амплитуду колебаний достаточно малой. Выбрав в качестве переменной заряд q на конденсаторе, с помощью уравнений Кирхгофа получим

$$\ddot{q} + (LC)^{-1} q = 0$$

Помножив обе части уравнения на $L\dot{q}$, получим

$$L\dot{q}\frac{d\dot{q}}{dt} + \frac{1}{C} q \frac{dq}{dt} = \frac{L}{2}\frac{d\dot{q}^2}{dt} + \frac{1}{2C}\frac{dq^2}{dt} =$$

$$= \frac{d}{dt}\left(\frac{L\dot{q}^2}{2} + \frac{q^2}{2C}\right) = \frac{d}{dt}(E_l + E_c) = 0$$

то есть для любого момента времени выполняется равенство $E_l + E_c = const$, отражающие постоянство во времени полной энергии осциллятора (суммы магнитной E_L и электрической E_C энергий).

Введем замену времени $\tau = \frac{t}{\sqrt{LC}}$, $dt = \sqrt{LC}d\tau$

$$\frac{L\dot{q}^2}{2} + \frac{q^2}{2C} = \frac{\dot{q}^2}{2C} + \frac{q^2}{2C} = a_0$$

Или, обозначая для общности q через x: $\dot{x}^2 + x^2 = a^2$, $a = const$.

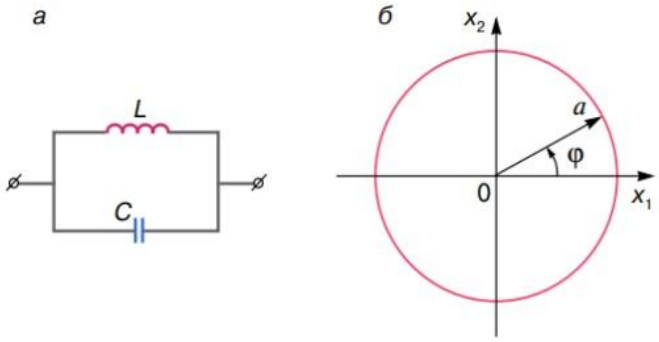


Рис. 1. а – колебательный контур, моделируемый в задаче; б – фазовый портрет колебаний при заданном уровне энергии

Для фазовых координат $x_1 = x$, $x_2 = \dot{x}$ эти уравнения преобразуются к виду

$\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1$, $x_1^2 + x_2^2 = a^2$.

Фазовый портрет системы представляет собой окружность радиуса *a* с центром в начале координат. Точка в фазовом пространстве, в которой вектор фазовой скорости обращается в нуль, называется особой, и в данном случае нуль координат есть *особая точка типа центр*.

Наличие интеграла движения у рассматриваемой системы, отражающее факт сохранения энергии, дает возможность описать ее с помощью уравнения 1-го порядка. Действительно, определив новую переменную φ соотношениями

$x_1 = a\sin\varphi$, $x_2 = b\cos\varphi$

получим уравнения $\dot{\varphi} = 1$, $\dot{a} = 0$ которые и представляют закон движения фазовой точки. Во времени эволюционирует одна переменная φ, и фазовое пространство консервативного осциллятора, таким образом, одномерно.

Гармоническим колебаниям осциллятора отвечает равномерное движение изображающей точки по окружности радиуса a, как это показано на рис. 1б.

Если *консервативная система нелинейна*, то ее фазовый портрет усложняется. Проиллюстрируем это на примере уравнения $\ddot{x} + \sin x = 0$.

В фазовых переменных $x_1 = x$, $x_2 = \dot{x}$ это уравнение сводится к следующим:

$\dot{x}_1 = x_2$, $\dot{x}_2 = -\sin(x)$.

Состояния равновесия нелинейного маятника на фазовой плоскости расположены вдоль оси x_1 ($x_2 = 0$) в точках $x_1 = 0, \pm\pi, \pm 2\pi, \ldots$. Соответствующий фазовый портрет системы представлен на рис. 2.

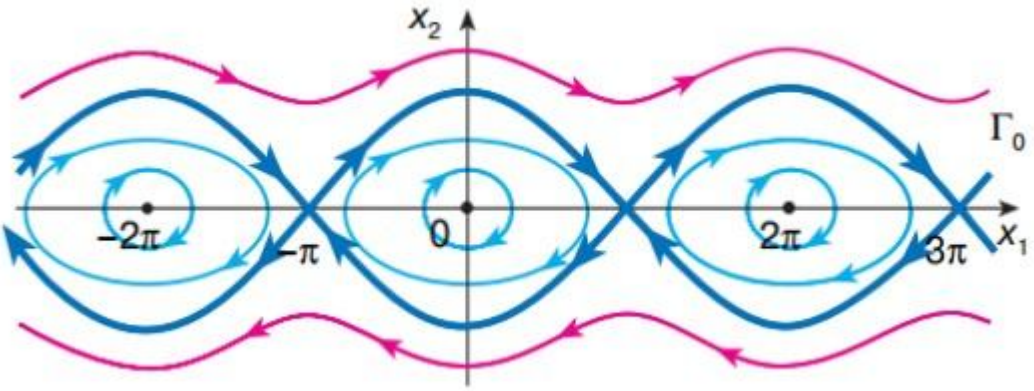


Рис. 2. Фазовый портрет осциллятора

Видно, что особые точки $x_1 = 0, \pm 2\pi, \pm 4\pi, \ldots$ типа *центр*, а $x_1 = \pm\pi, \pm 3\pi, \ldots$ – неустойчивые точки типа *седло*.

Вблизи центров фазовый портрет соответствует линейному осциллятору: траектории представляют собой замкнутые кривые, близкие к окружностям, что отвечает по амплитуде колебаниям, близким к гармоническим.

Через неустойчивые точки проходят особые интегральные кривые Γ_0, называемые *сепаратрисами.* Они разделяют фазовое пространство на области с различным поведением. С увеличением энергии маятника его колебания от квазигармонических вблизи точек типа центр эволюционируют к нелинейным периодическим колебаниям вблизи сепаратрис.

Дальнейшее увеличение энергии приведет к *вращательному движению* (движение вне сепаратрис). Малейшие отклонения энергии в ту или иную сторону от энергии движения по сепаратрисе приводят к качественно различным типам движения: колебательному или вращательному.

Линейный осциллятор с затуханием. Диссипация энергии, обусловленная наличием потерь, оказывает принципиальное влияние на характер движения системы. Наиболее простые закономерности проявляются в системах с полной диссипацией энергии, когда силы трения действуют по всем степеням свободы, а поступление энергии извне отсутствует.

Рассмотрим процессы в линейном диссипативном осцилляторе, когда сила трения пропорциональна скорости изменения координаты. Примером такой системы служит колебательный контур, содержащий активное сопротивление R. Уравнение контура

$$L\ddot{q} + R\dot{q} + \frac{q}{C} = 0$$

заменой переменных сводится к безразмерной форме

$$\ddot{x} + 2\delta\dot{x} + x = 0, \quad 2\delta = R\sqrt{\frac{L}{C}}, \quad \tau = \frac{1}{\sqrt{LC}}.$$

При $\delta = 0$ имеем консервативный линейный осциллятор, рассмотренный выше. Введение малого трения качественно меняет фазовый портрет системы.

Для $0 < \delta < 1$ решением последнего уравнения (20) является

$$x = A\exp(-\delta\tau)\cos(\omega t + \psi), \quad \omega = \sqrt{1-\delta^2}\,.$$

где A и ψ – произвольные постоянные, определяемые начальными условиями.

На фазовой плоскости для любых начальных данных имеют место скручивающиеся спирали, по которым фазовые точки асимптотически приближаются к началу координат, характеризуя затухающий колебательный процесс.

Нуль координат является особой точкой системы, которая в случае δ < 1 есть *устойчивый фокус* (рис. 3а).

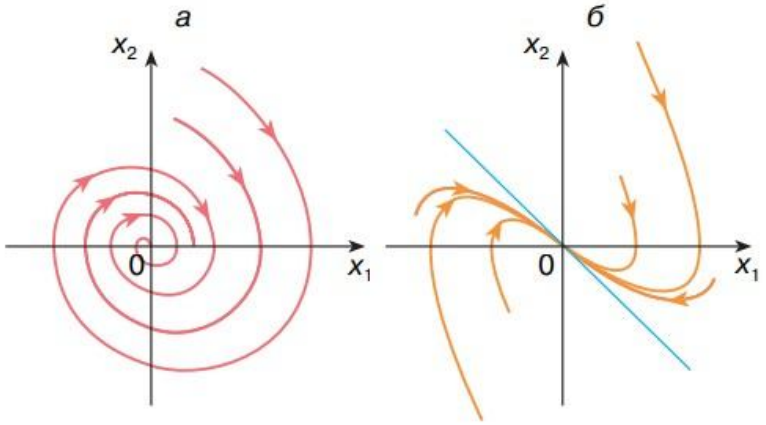


Рис. 3. Фазовый портрет диссипативного осциллятора с параметром δ < 1 (а) и δ > 1

Если коэффициент трения δ>1, процесс в системе апериодический:

$x = A_1 \exp(\lambda_1 \tau) + A_2 \exp(\lambda_2 \tau)$, $\lambda_{1,2} = [-\delta \pm (\delta_2 - 1)^{1/2}]/2$

и фазовые траектории выглядят как семейство характерных кривых, по которым, как и в предыдущем случае, изображающие точки стремятся к нулю координат (рис. 3б).

Особая точка в указанных условиях является *устойчивым узлом*.

Итак, при любых значениях физических параметров системы, когда δ > 0, диссипативный маятник характеризуется единственным глобально устойчивым состоянием равновесия в нуле фазовых координат. Независимо от выбора начальных условий наблюдается затухающее колебательное или апериодическое движение.

При $t \to \infty$ любая изображающая точка стремится к началу координат либо в устойчивый фокус, либо в узел.

Описанное свойство является общим для *динамических систем с полной диссипацией энергии*. Положения равновесия типа устойчивого фокуса или узла являются здесь глобально притягивающими в том смысле, что фазовые траектории из любой точки фазового пространства асимптотически к ним стремятся.

Стационарные незатухающие колебания в линейных диссипативных системах оказываются невозможными. С физической точки зрения это понятно – нет условий поддержания колебаний. Энергия, расходуемая на преодоление сил трения, не восполняется.

6. Автоколебательные системы

Возможность существования периодического асимптотически устойчивого движения, изображаемого изолированной замкнутой траекторией в фазовом пространстве, к которой со временем притягиваются траектории из некоторой окрестности независимо от начальных условий, обеспечивается только в нелинейных диссипативных системах. Этот тип динамических систем настолько важен при

изучении колебательных процессов, что для его выделения А.А. Андронов предложил специальный термин – *автоколебательные системы*.

Математическим образом автоколебаний служит *предельный цикл Пуанкаре* – замкнутая изолированная траектория в фазовом пространстве, отвечающая периодическому движению.

В качестве примера ДС с предельным циклом Пуанкаре рассмотрим классический *нелинейный осциллятор Ван дер Поля*, уравнение колебаний которого

$$\ddot{x} - a(1 - bx^2)\dot{x} + x = 0$$

Параметр ***a***, характеризующий подкачку энергии в систему от внешнего источника, является существенным параметром осциллятора и называется *параметром возбуждения*. Из сравнения уравнений Ван дер Поля и диссипативного осциллятора следует, что осциллятор Ван дер Поля описывает более сложный колебательный контур, характер диссипации в котором зависит от переменной ***x***.

В фазовых координатах уравнение колебаний осциллятора Ван дер Поля представляется как

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = a(1 - bx_1^2)x_2 - x_1,$$

причем $a(1 - bx_1^2) \neq 0$.

Аналитически представленные уравнения не решаются, и исследования проводятся с использованием численных методов.

В практически важном случае (a>0, b>0) данные уравнения имеют единственное устойчивое решение в виде *предельного цикла* Г, изображенного на рис. 4а.

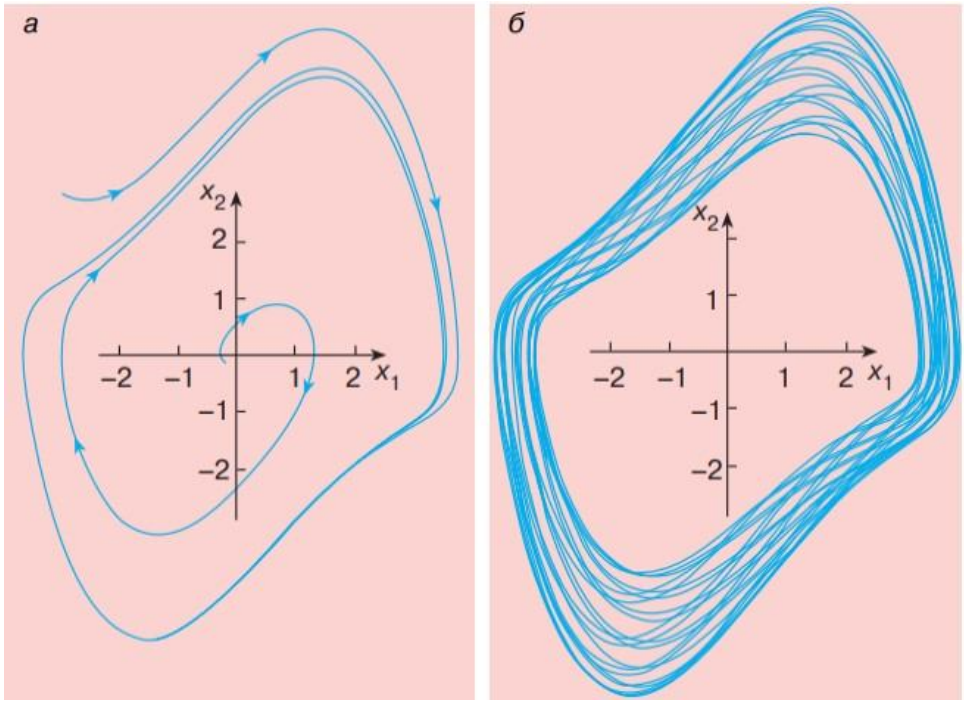


Рис. 4. Предельный цикл системы; расчет для значений параметров a = 1, b = 0,3 (а).
Проекция двумерного тора на плоскость переменных x1, x2 ; численное интегрирование уравнений для значений параметров a = 1, b = 0,3, B = 1,0, ϕ0 = 0 (б)

Положение в начале координат, в котором вблизи нуля можно пренебречь нелинейностью, является неустойчивым фокусом. Траектории из окрестности состояния равновесия асимптотически стремятся к предельному циклу. Как показывает анализ, предельный цикл является устойчивой изолированной структурой, притягивающей к себе траектории из любой точки на фазовой плоскости.

Таким образом, в динамических системах с нелинейной зависимостью диссипации энергии от переменной, совершающей колебания, впервые появляется принципиально новый тип устойчивого предельного множества фазовых траекторий – предельный цикл. На предельном цикле за время периода колебаний доли рассеиваемой и вносимой энергии строго компенсируются.

Наконец, рассмотрим еще один случай типичной структуры в фазовом пространстве ДС, возникающей, например, при периодическом возмущении системы с устойчивым предельным циклом.

Добавим в уравнение Ван дер Поля источник гармонического действия сравнительно малой амплитуды B и частоты p, которую считаем рационально не связанной с частотой периодических колебаний автономного осциллятора:

$$\ddot{x} - a(1 - bx^2)\dot{x} + x = B\sin(p\tau + \varphi_0).$$

Периодическая модуляция предельного цикла автономной системы приводит к тому, что фазовая траектория с заданной частотой p вращается вокруг предельного цикла и лежит на двумерной поверхности, представляющей собой поверхность тора. Аналогично случаю предельного цикла эта поверхность будет устойчивым предельным множеством, к которому стягиваются со временем все траектории из некоторой окрестности тора (как изнутри него, так и снаружи). Нетрудно представить себе, что минимальная размерность фазового пространства, в которое можно вложить двумерный тор, равна трем.

На рис. 4б показана проекция на плоскость переменных x_1, x_2 фазовой траектории на двумерном торе, полученная численным интегрированием системы уравнений колебаний осциллятора Ван дер Поля.

7. Регулярные и странные аттракторы

Рассмотренные примеры иллюстрируют типичные предельные множества траекторий на фазовой плоскости: состояния равновесия, периодические движения и особые траектории типа сепаратрисных контуров.

Указанные предельные множества полностью исчерпывают возможные ситуации на фазовой плоскости. Им отвечают три различных типа решений уравнений. Движения диссипативных систем целесообразно разделить на два класса: класс переходных, нестационарных движений, отвечающих переходу от начального к предельному множеству состояний, и класс установившихся стационарных движений, фазовые траектории которых целиком принадлежат предельным множествам.

Важными с физической точки зрения являются притягивающие предельные множества – *аттракторы*. С течением времени произвольное начальное состояние из некоторой области притяжения G, включающей в себя аттрактор G0, переходит к G0. Движение, которому отвечает фазовая траектория в области притяжения, есть переходной процесс. Установившееся движение характеризуется принадлежностью фазовых траекторий предельному множеству, то есть аттрактору G0.

К чему может привести повышение размерности системы, например, до N = 3, то есть выход с плоскости в трехмерное фазовое пространство? Совсем недавно, до начала 60-х годов, с увеличением размерности фазового пространства дис-

сипативных систем связывали возможность появления (в дополнение к указанным выше) лишь квазипериодических аттракторов, соответствующих движениям на r-мерных торах.

Важным результатом исследований последующих лет явилось обнаружение принципиально новых типов движений в ДС. Таким движениям в фазовом пространстве размерности N≥3 соответствуют сложным образом устроенные притягивающие множества, траектории изображающих точек которых не принадлежат ни к одному из описанных выше типов аттракторов. Фазовые траектории представляются здесь в виде бесконечной, нигде не пересекающейся линии. При $t \to \infty$ траектория не покидает замкнутой области и не притягивается к известным типам аттракторов [2–6]. Именно с существованием таких траекторий связывают возможность *хаотического поведения детерминированных динамических систем с размерностью фазового пространства N≥3.*

Впервые подобные свойства ДС в 1963 году обнаружил Э.Лоренц при численном исследовании динамики трехмерной модели тепловой конвекции.

Спустя восемь лет в теоретической работе Д. Рюэля и Ф. Такенса притягивающая область в фазовом пространстве ДС, характеризуемая режимом установившихся непериодических колебаний, была названа *странным аттрактором*. Этот термин был сразу воспринят исследователями и утвердился для обозначения математического образа режима нерегулярных колебаний детерминированных динамических систем [2–6].

Аттракторы в виде состояний равновесия, предельных циклов или l-мерных торов называют *простыми или регулярными*, подчеркивая тем самым, что движения на них отвечают сложившимся представлениям об устойчивом по Ляпунову детерминированном поведении ДС.

Со странным аттрактором связывается реализация *нерегулярного (в смысле отсутствия периодичности) колебательного режима*, который во многом сходен с нашими представлениями о стационарных случайных процессах. Термин случайный имеет вполне определенный смысл.

Случайное движение непредсказуемо либо предсказуемо с определенной вероятностью. Другими словами, траектории случайного движения нельзя многократно и однозначно воспроизвести ни в численном, ни в физическом эксперименте. Примером служит классическое движение броуновской частицы.

В случае странного аттрактора имеется строгая предсказуемость в смысле детерминированности закона эволюции. Решение уравнений (как и для регулярных аттракторов) подчиняется теореме единственности и однозначно воспроизводится при фиксированных начальных условиях. Поэтому для обозначения сложных “шумоподобных” автоколебаний, математическим образом которых служит странный аттрактор, используются термины типа *динамическая стохастичность, детерминированный хаос* и подобные.

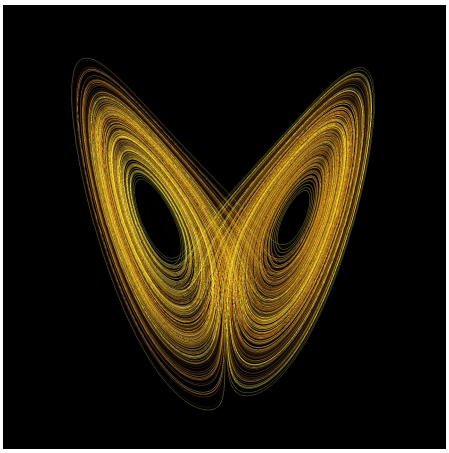


Рис. 5. Вид аттрактора Лоренца

Примером странного аттрактором является аттрактор Лоренца, определяемый системой уравнений

$$\dot{x} = \sigma(y - x); \quad \dot{y} = x(r - z) - y; \quad \dot{z} = xy - bz$$

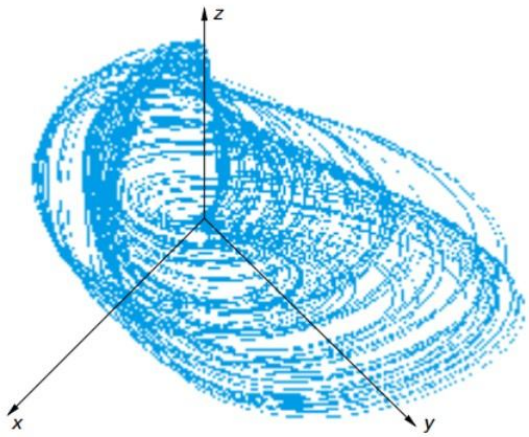
при следующих значениях параметров: σ=10, *r*=28, *b*=8/3, x(0)=1, y(0)=0, z(0)=0.

Важно отличать эти процессы от стохастических в классическом смысле, которые при описании требуют учета флуктуаций в исходных динамических уравнениях либо непосредственно подчиняются уравнениям для плотности распределения вероятностей статистической теории [2, 5].

Примером системы с хаотическим аттрактором являются уравнения генератора с инерционной нелинейностью (генератора Анищенко–Астахова). Эта система является обобщением уравнений Ван дер Поля на случай трехмерного пространства [2]:

$$\dot{x} = mx + y - xz; \quad \dot{y} = -x; \quad \dot{z} = -gz + gI(x)x^2,$$

$$I(x) = \begin{cases} 1, & x > 0, \\ 0, & x <= 0. \end{cases}$$



Результаты численного решения уравнения этой системы уравнений для значений параметров m = 1,5, g = 0,2 приведены на рис. 5, который также иллюстрирует хаотический аттрактор.

Рис. 5. Странный аттрактор в модели генератора Анищенко–Астахова

Выводы

Дано общее определение ДС и приведены примеры динамических систем, описываемых обыкновенными дифференциальными уравнениями. Такие ДС могут иметь четыре типа решений:

- состояние равновесия,
- периодическое движение,
- квазипериодическое движение,
- хаотическое.

Этим типам решений соответствуют аттракторы системы в виде устойчивого равновесия, предельного цикла, квазипериодического аттрактора (p-мерного тора) и хаотического (или странного) аттрактора. Важным является то, что простейшие типы квазипериодических и хаотических аттракторов могут реализовываться в ДС с размерностью фазового пространства не менее трех.

Вопросы для самопроверки:

1.Что называется ДС?
2. Что означает задание математической модели ДС?
3. Приведите пример нелинейного консервативного осциллятора.
4. Поясните природу множественности математических моделей.
5. Что называется фазовой точкой? Фазовым пространством?
6. Приведите векторную форму записи ДС.
7. Что называется числом степеней свободы?
8. Что называется оператором отображения?

9. Запишите свойство суперпозиции для линейных систем.
10. Что называется потоками? Каскадами?
11. Какие системы называются нелинейными? Пример.
12. Какие системы называются сосредоточенными? Распределенными?
13. Какие системы называются гамильтоновыми?
14. Напишите уравнение линейного консервативного осциллятора.
15. Напишите уравнение линейного осциллятора с затуханием.
16. Приведите пример нелинейного консервативного осциллятора.
17. Какие системы называются автоколебательными?
18. Что называется аттрактором? Странным аттрактором?

Литература:

1. Аносов Д.В. ДС // Математическая энциклопедия. М.: Сов. энциклопедия, 1979.
2. Анищенко В.С. Сложные колебания в простых системах. М.: Наука, 1990.
3. Лихтенберг А., Либерман М. Регулярная и стохастическая динамика. М.: Мир, 1984.
4. Шустер Г. Детерминированный хаос. М.: Мир, 1988.
5. Неймарк Ю.И., Ланда П.С. Стохастические и хаотические колебания. М.: Наука, 1987.
6. Лоскутов А.Ю., Михайлов А.С. Введение в синергетику. М.: Наука, 1990.
7. Анищенко В.С. ДС. – URL: http://www.pereplet.ru/nauka/Soros /pdf/ 9711_077.pdf
8. Юмагулов М.Г. Введение в теорию динамических систем: Учебное пособие. – М.: Лань. 2015. -272с.
9. Неймарк Ю.И. ДС и управляемые процессы. – М.: Либроком. 2010. – 338с.
10. А. Каток, Б. Хасселблат. Введение в современную теорию динамических систем. М.: Факториал, 1999.

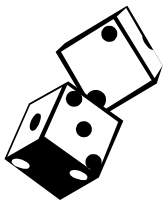
Приложение 1. Программа генерации аттрактора Лоренца.

```
%Solution for the Lorenz equations in the time interval [0,100] with initial conditions [1,1,1].

clear all; clc;
sigma=10; beta=8/3;
rho=28;
f = @(t, a) [-sigma*a(1) + sigma*a(2); rho*a(1) - a(2) - a(1)*a(3); -beta*a(3) + a(1)*a(2)];
 [t, a] = ode45(f, [0 100], [1 1 1]);       %'ode45' uses adaptive Runge-Kutta method of 4th
                                            and          % 5th order to solve differential equations
plot3(a(:,1),a(:,2),a(:,3))      %'plot3' is the command to make 3D plot
```

ЛЕКЦИЯ 6

ПРОБЛЕМА НЕОПРЕДЕЛЕННОСТИ



Мы живем в вероятностном мире. Осознание этого факта потребовало от человечества пройти долгой дорогой исканий от аристотелевского детерминизма до принципа неопределенности Гейзенберга и случайной вселенной Паули-Юнга, от философских размышлений в садах Академа и Ликея до современных систем статистического анализа данных, упакованных в компакты программных продуктов.

Что же представляет из себя случайность?

Жизнь сталкивает человека с неопределенностью на каждом шагу. Все наше будущее в той или иной форме случайно. Прошлое, напротив, строго детерминировано, как детерминирована любая конкретная реализация.

С точки зрения систем принятия решений, вероятностное будущее, ближайшее и отдаленное, имеет несравненно большее значение, чем уже определившееся, навсегда детерминированное прошедшее. Однако для предсказания будущего, определяющего качество формируемых управляющих решений, нет иного источника, кроме прошлого, точнее, кроме накопленных в прошлом знаний. И уж конечно, вся научная деятельность человека обращена сугубо в его вероятностное будущее.

1. Введение. Детерминизм. Закономерность против случайности

Со времен античности *закономерности* (предопределенность) окружающего нас мира и человеческой жизни соотносилась с порядком, с тем, что можно предсказать, а *случайность* рассматривалась как отклонение от порядка, нарушение порядка, погружение в хаос.

Иными словами, механизм, определяющий существование закономерности, опирается на причинно-следственные связи, которые являются ключевым фактором всех событий и явлений.

Причина порождает следствие, а следствие является новой доминантой до тех пор, пока не приобретает качество причины для следующего этапа.

Существует гипотеза о предопределенности будущего, описываемой жесткой схемой причинно-следственных связей или божественным промыслом.

Альтернативная гипотеза состоит в *многовариантной природе будущего*, на которое можно влиять путем принятия тех или иных субъективных решений. При этом формирование решений, в том числе и научных, осуществляется *в условиях неопределенности*, характеризуемой отсутствием достаточных знаний, как об окружающей среде, так и о внутренних (по отношению к анализируемой системе) процессах.

Именно отсутствие необходимого объема знаний провозглашалось основным и единственным генезисом случайности. Однако недостаток знаний — это лишь субъективный фактор, связанный с пониманием человеком окружающей его действительности. Если же изменения в природе определяются четкими причинно-следственными связями, значит все в ней детерминировано, никакой неопределенности, а, следовательно, и случайности не существует.

Вывод детерминистов достаточно очевиден**:** понятие случайности является субъективным и существует лишь в сознании человека, недостаточно образованного и недостаточно понимающего окружающий однозначный мир.

Этой же, детерминистической точки зрения придерживался Фридрих Великий: «Нет ничего случайного. Случайность – это то, что недоступно видению».

Принцип детерминизма, сформулированный *П.-С. Лапласом* в его сочинении "Опыт философии теории вероятностей", утверждает, что при обладании необходимым объемом достоверной информации любое событие в будущем является абсолютно прогнозируемым.

Так, например, результат пресловутого подбрасывания монетки, превратившегося в символ генерации событий с 50%-ной вероятностью реализации, вполне может быть предсказан, если точно знать величину полученного механического импульса, расположение точки его приложения относительно центра масс, высоту руки над поверхностью падения, характеристики твердости и гладкости поверхности и многое, многое другое.

Пьер-Симон Лаплас *(1749, Кальвадос —1827, Париж) — выдающийся французский математик, физик и астроном; известен работами в области небесной механики, дифференциальных уравнений, один из создателей теории вероятностей. Заслуги Лапласа в области чистой и прикладной математики и особенно в астрономии громадны: он усовершенствовал почти все отделы этих наук. Был членом АН и Французского Географического общества.*

Демон Лапласа – великий всезнайка, способный сформировать достоверный прогноз, точно зная всю совокупность факторов влияния, вплоть до динамики микрочастиц.

Следует заметить, что идея детерминизма была провозглашена философами задолго до Лапласа и Фридриха. Так, Аристотель писал: *«Ничто не делается случайно. Для всего, возникновению чего мы приписываем самопроизвольности или случаю, имеется некоторая определенная причина».* Ему же принадлежит фраза: *«Нет ничего более противного разуму и природе, чем случайность*».

Однако в том и заключалось диалектическое величие древнегреческих мыслителей, что они умели для каждого мудрого тезиса находить и обосновывать не менее убедительный антитезис. Говоря словами Протагора, *«каждому рассуждению противостоит равносильное».*

И вот он, антитезис Ксенофана: "Нет, достоверно никто никогда ничего не узнает".

2. Вероятностный мир: история развития. Этот случайный мир

Тем не менее, случайность реально существует и окружает нас непрерывно как в науке, так и в повседневной жизни.

Указывая на всеобщность вероятностного подхода, римский поэт Публий Сир писал: *"В каждом большом деле всегда приходится какую-то часть оставить на долю случая"*.

В человеческом сознании случайность ассоциируется с хаосом, активно противодействующим благим намерениям *homo sapiens*. Вселенский закон бутерброда ("Бутерброд всегда падает маслом вниз") уже в наше время дополнен целым рядом выстраданных учеными положений "мэрфологии" [4].

Могучий аппарат теории вероятностей и математической статистики позволяет существенно облегчить общение со случайностью, пронизывающей всю нашу жизнь. Вероятность в научных исследованиях - категория весьма непростая, и прежде чем вступить в ее стохастические дебри, не лишне взглянуть на них сверху и наметить дальнейший путь.

Закон Мэрфи (основной закон мэрфологии): если какая-нибудь неприятность может произойти, она случается.

Первый закон Чизхолма: Все, что может испортиться, портится.

Следствие: Все, что не может испортиться, портится тоже.

Закон своенравия природы: нельзя заранее правильно определить, какую сторону бутерброда мазать маслом.

Принцип очереди: чем больше ожидание, тем больше вероятность, что вы стоите не в той очереди.

Идея лапласовского детерминизма довольно долго властвовала в науке, не имея достаточно убедительных альтернатив в виде источников тотальной, ничем не обусловленной случайности. И лишь поистине величайшие открытия физики XX столетия в мире микрочастиц позволили отыскать источники "первозданной" неопределенности. Впрочем, в мире, где разрушены даже причинно-следственные связи, можно, при желании, найти все, что угодно. В том числе - и самого Нечистого, творца Хаоса.

Следует заметить, что математики не отстали от своих физических "братьев по разуму" и отыскали "беспричинно" случайные процессы при решении некоторых вполне детерминированных нелинейных дифференциальных уравнений [2].

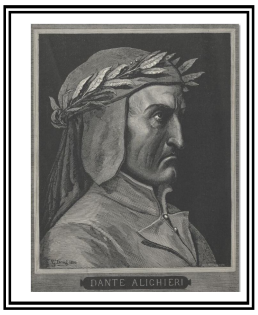
Разумеется, для большинства прикладных наук вопрос о генезисе случайности является вторичным. Важно другое - научиться в этом случайном мире формировать разумные (насколько это возможно) управляющие решения. При этом решения должны базироваться на современной научной методологии, прежде всего, математической.

Проблема оказалась настолько важной, что, по мнению Н. Винера, обобщается на всю формализованную методологию: *"... высшее назначение математики как раз и состоит в том, чтобы находить скрытый порядок в хаосе, который нас окружает".* При этом современная теория вероятностей дает не только технологию статистических вычислений, но и некоторую фундаментальную концепцию, позволяющую найти порядок и закономерность там, где классический детерминистический подход оказывается бессильным. Она предлагает более широкое понимание причинных связей, чем это делает любая детерминистическая теория.

Как отмечалось выше, интерес к случайности обнаруживается уже в трудах великих мыслителей древнего мира - Демокрита, Платона, Аристотеля. Известно, что статистические выводы общего характера делались в древнем Китае еще в

2238г. до н.э., в эпоху императора Яо на основе анализа результатов переписи населения в Поднебесной. Однако мысль о возможности характеризовать числом и мерой степень случайности возникла гораздо позднее, в районе XVI-XVII вв.

С диалектической терпимостью отметим, что азартные игры (от французского слова "hasard", буквально означающего "случай", "риск") послужили мощным стимулятором развития вероятностной науки. В своем письме к Ф. ван Схоутену (1657), автору книги «О расчетах в азартных играх", Христиан Гюйгенс отмечает: "...я полагаю, что при внимательном изучении предмета читатель заметит, что имеет дело не только с игрой, но что здесь закладываются основы очень интересной и глубокой теории».



Значительно ранее (период кватроченто), в «Божественной комедии» Данте Алигьере указывается на попытки подсчитать число благоприятных исходов при игре в кости.

Схема азартных игр обладает кажущейся простотой и доступностью для формальной логики. Первые попытки этого рода связаны с именами известных учёных - алгебраиста *Джироламо Кардана* (1501- 1576) и *Галилео Галилея* (1564-1642).

Основателями теории вероятностей, по-видимому, следует считать замечательных французских математиков XVII в. Б.Паскаля и П.Ферма. Инициацией исследований для Паскаля послужила *задача кавалера де Мере*.

Вариант вопроса состояла в следующем: как «по-справедливости» разделить начальные ставки между игроками (имеющими разное количество выигранных партий), закончившими игру до завершения заранее оговоренного общего числа партий. Другой вопрос кавалера де Мере заключался в следующем: что более вероятно при четырехразовом бросании кости – выпадение шестерки хотя бы один раз или ни разу?

Примечание. Вообще, игроки в азартные игры не производили подсчетов разных комбинаций очень долго. Это вызвано в первую очередь тем, что игроки в процессе игры менялись ролями, и не равносильность исходов каждому игроку была поочередно то выгодной, то невыгодной. Подсчёты вероятностей в азартных играх начались с того времени, когда в игре появилось два лагеря: козлик и игроки. Их роли в игре были разными. Козлику игры нужно было строить её ход таким образом, чтобы деньги между играющими перераспределялись (чтобы были выигравшие и проигравшие) и чтобы определённый процент денег играющих переходил неизменно к нему. Когда возникла такая ситуация, тогда и возник подсчёт вероятностей.

Наряду с вероятностными методами в трудах ученых-демографов XVII в. Д. Граунта и У. Петти закладывались начала математической статистики.

Тотальную значимость вероятностной концепции достаточно быстро поняли и философы. Так, уже в 1687 г. был опубликован трактат Б. Спинозы «*Заметки о математической вероятности».*

Дальнейшее развитие теории вероятностей и математической статистики связано с именами Я. Бернулли, А. де Муавра, Монмора, Н. Бернулли, Д. Бернулли, Л. Эйлера, Т. Байеса, Ж. Даламбера и других корифеев мировой математики. Русская школа теории вероятностей хорошо известна трудами П. Л. Чебышева, М. В. Остроградского, В. Я. Буняковского, А. Н. Колмогорова и других.

Впервые курс теории вероятностей в России был введен решением Совета Михайловской артиллерийской академии в Санкт-Петербурге в 1858 г. и прочитан ее слушателям профессором М. В. Остроградским. Впрочем, это далеко не един-

ственный случай, когда потребности военной науки стимулировали развитие и прогресс вероятностной методологии.

3. Начала теории вероятностей

Теория вероятностей — раздел математики, изучающий закономерности случайных явлений: случайные события, случайные величины, их свойства и операции над ними.

Вероятность (вероятностная мера) — численная мера возможности наступления некоторого события.

С практической точки зрения, вероятность события — это отношение количества тех наблюдений, при которых рассматриваемое событие наступило, к общему количеству наблюдений. Такая трактовка допустима в случае достаточно большого количества наблюдений или опытов.

Например, если среди встреченных на улице людей примерно половина — женщины, то можно говорить, что вероятность того, что встреченный на улице человек окажется женщиной, равна 1/2. Другими словами, оценкой вероятности события может служить частота его наступления в длительной серии независимых повторений случайного эксперимента.

Если каждому элементарному событию поставить в соответствие число $p_i \in [0,1]$, для которого выполняется условие $\sum_{i=1}^{n} p_i = 1$, то считается, что заданы вероятности элементарных событий.

Вероятность события, как счётного подмножества пространства элементарных событий, определяется как сумма вероятностей тех элементарных событий, которые принадлежат этому событию. Требование счётности существенно, так как, иначе сумма будет не определена.

Рассмотрим правила определения вероятности различных случайных событий. Например, если событие является пустым множеством Λ, то его вероятность равна нулю: $P(\Lambda) = 0.$

Если событие совпадает со всем пространство элементарных событий Ω, то его вероятность равна единице: $P(\Omega) = 1.$

Вероятность события (подмножества пространства элементарных событий) равна сумме вероятностей тех элементарных событий, которые включает в себя рассматриваемое событие.

Оценка вероятности возникновения события может быть определена через частоту этого события

$$f_n = \frac{m}{n},$$

где m – число реализаций изучаемого события в **n** опытах, повторенных с неизменных условия. Тогда вероятность определяется как $p = \lim(f_n) \quad \text{при} \quad n \to \infty.$

Рассмотрим задачу построения так называемой функции распределения вероятностей случайной величины x *F(X) = P(x<=X)* на примере бросания игральной кости.

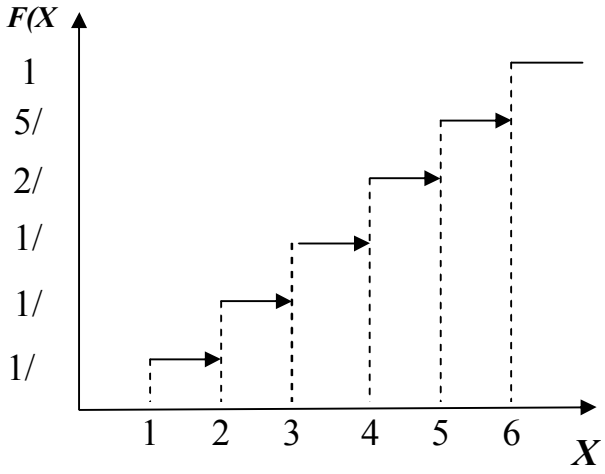


Рис. 1

Множество возможных исходов *X={0, 1, 2, 3, 4, 5, 6}.*

При этом соответствующие вероятности исходов равны

P(0)=(x<0)=0;
P(1)=P(x<=1)=P(x=1)=1/6;
P(2)=P(x<=2)=P(x=1Vx=2)= =2/6=1/2;
.....
P(6)=P(x<=6)=1.

Графически эта функция будет иметь вид, представленный на рис. 1.

Предположим теперь, что у нас 10 000 граней у гиперкубика, на которых с равномерным шагом 1/10000 нанесены цифры в интервале от -5 до 5.

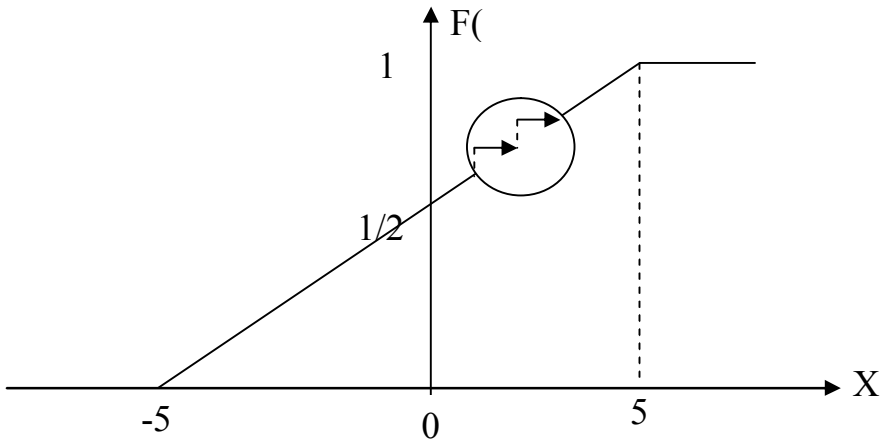


Рис. 2

тогда получим картинку распределения, близкую к непрерывному распределению вида, представленного на рис. 2. Такая функция распределения называется равномерной.

Производная от функции распределения дает возможность судить о скорости изменения вероятности события в заданном интервале значений случайной величины и называется плотностью распределения или дифференциальной функций распределения. Пример плотности равномерного распределения вероятностей приведен на рис. 3.

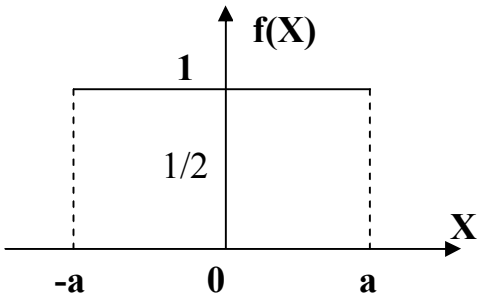


Рис. 3

Самым распространенным типом функции распределения является нормальное распределение, также называемое гауссовским распределением или распределением Гаусса.

Данное распределение вероятностей, которое задается функцией плотности распределения: $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}$, где параметр μ — среднее значение (математическое ожидание) случайной величины и указывает координату максимума кривой плотности распределения, а σ^2 — дисперсия.

Вид нормальной функции распределения и ее плотности приведены на рис.4.

Нормальное распределение играет важнейшую роль во многих областях знаний, особенно в статистической физике.

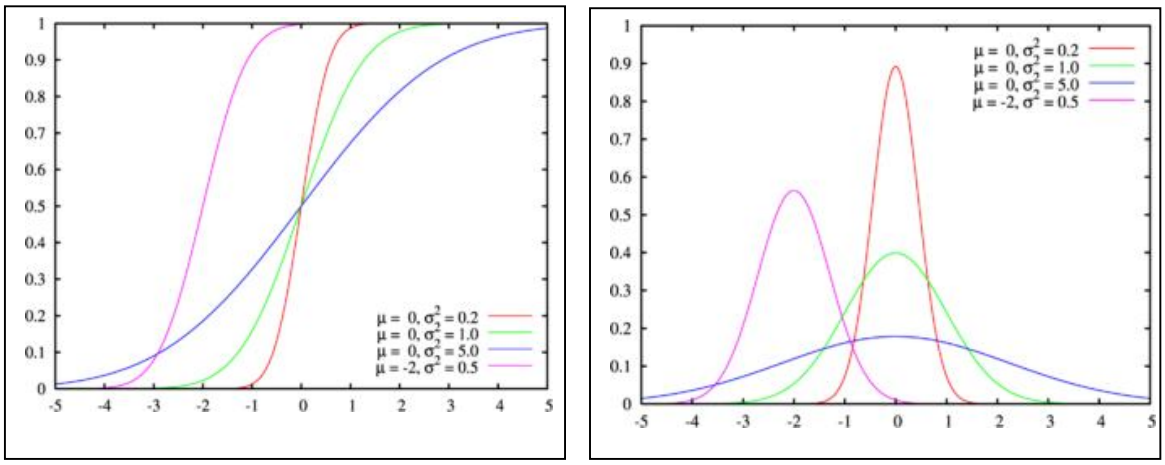


Рис. 4. Функция и плотность нормального распределения для различных значений математического ожидания и дисперсии

Физическая величина, подверженная влиянию значительного числа независимых факторов, могущих вносить с равной погрешностью положительные и отрицательные отклонения, вне зависимости от природы этих случайных факторов, часто подчиняется нормальному распределению, поэтому из всех распределений в природе чаще всего встречается нормальное (отсюда и произошло это название этого распределения вероятностей).

Иоганн Карл Фридрих Гаусс (1777, Брауншвейг —1855, Гёттинген) — немецкий математик, астроном и физик, считается одним из величайших математиков всех времён, «королём математиков».

Нормальное распределение зависит от двух параметров — *смещения* и *масштаба*, то есть является с математической точки зрения не одним распределением, а целым их семейством. Значения параметров соответствуют значениям среднего (математического ожидания) и разброса (стандартного отклонения).

Стандартным нормальным распределением называется нормальное распределение с математическим ожиданием *0* и стандартным отклонением *1.*

Математическое ожидание — числовая характеристика параметра положения распределения вероятностей случайной величины. Представляет собой взвешенное среднее возможных значений случайной величины.

В англоязычной литературе и в математическом сообществе Санкт-Петербурга обозначается через *E{X}* (например, от англ. *Expected value*), в русской и московской научных школах — *M{X}* (возможно, от англ. *Mean value*, а возможно от рус. *Математическое ожидание*). В статистике часто используют обозначение μ.

Среднеквадратическое отклонение (синонимы: среднеквадратичное отклонение, квадратичное отклонение; близкие (но не совпадающие) термины: стан-

дартное отклонение, стандартный разброс) — в теории вероятностей и статистике наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания.

На рис. 5 проиллюстрирован смысл этих понятий.

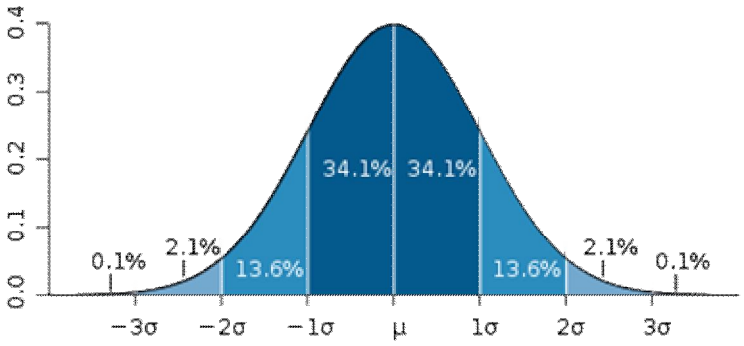


Рис. 5. Иллюстрация понятия ско

4. Предельные теоремы

Важнейшими теоремами теории вероятностей являются закон больших чисел и предельные теоремы.

Закон больших чисел в теории вероятностей утверждает, что эмпирическое среднее (среднее арифметическое) достаточно большой конечной выборки из фиксированного распределения близко к теоретическому среднему (математическому ожиданию) этого распределения. В зависимости от вида сходимости различают слабый закон больших чисел, когда имеет место сходимость по вероятности, и усиленный закон больших чисел, когда имеет место сходимость почти всюду.

Всегда найдётся такое количество испытаний, при котором с любой заданной наперёд вероятностью относительная частота появления некоторого события будет сколь угодно мало отличаться от его вероятности.

Общий смысл закона больших чисел — совместное действие большого числа случайных факторов приводит к результату, почти не зависящему от случая!

На этом свойстве основаны методы оценки вероятности на основе анализа конечной выборки.

Наглядным примером является прогноз результатов выборов на основе опроса выборки избирателей.

Усиленный закон больших чисел. Пусть есть бесконечная последовательность независимых одинаково распределённых случайных величин $\{X_i, \quad i = 1,...,\infty\}$, определённых на одном вероятностном пространстве (Ω, F, P). Пусть $EX_i = \mu, \forall i = 1,...,n.$. Обозначим S_n выборочное среднее первых n членов:

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad n \in N.$$

Тогда $S_n \underset{n\to\infty}{\to} \mu$ почти наверное.

Центральные предельные теоремы (Ц.П.Т.) — класс теорем в теории вероятностей, утверждающих, что сумма достаточно большого количества слабо зависимых случайных величин, имеющих примерно одинаковые масштабы (ни одно из слагаемых не доминирует, не вносит в сумму определяющего вклада), имеет распределение, близкое к нормальному.

Так как многие случайные величины в приложениях формируются под влиянием нескольких слабо зависимых случайных факторов, их распределение считают нормальным. При этом должно соблюдаться условие, что ни один из факторов не является доминирующим. *Центральные предельные теоремы* в этих случаях обосновывают применение нормального распределения.

Классическая формулировка центральной предельной теоремы (ЦПТ)

Пусть $X_1, X_2, ..., X_n, ...$ есть бесконечная последовательность независимых одинаково распределённых случайных величин, имеющих конечное математическое ожидание μ и дисперсию σ^2. Пусть $S_n = \sum_{i=1}^{n} X_i$. Тогда

$$P\{\frac{S_n - \mu n}{\sigma\sqrt{n}}\} \Rightarrow N(0,1) \quad \text{при } n \to \infty,$$

где *N(0, 1)* — нормальное распределение с нулевым математическим ожиданием и стандартным отклонением, равным единице.

Пусть $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ - выборочное среднее первых *n* величин, то есть, мы можем переписать результат центральной предельной теоремы в следующем виде:

$$P\{\sqrt{n}\frac{X - \mu}{\sigma}\} \Rightarrow N(0,1) \text{ при } n \to \infty.$$

Замечания

- Неформально говоря, классическая центральная предельная теорема утверждает, что сумма *n* независимых одинаково распределённых случайных величин имеет распределение, близкое к $N(n\mu, n\sigma^2)$.

Эквивалентно, $\overline{X}$ имеет распределение близкое к $N(\mu, \sigma^2/n)$.

- Так как функция распределения стандартного нормального распределения непрерывна, сходимость к этому распределению эквивалентна поточечной сходимости функций распределения к функции распределения стандартного нормального распределения. Положив $Z_n = \frac{S_n - \mu n}{\sigma\sqrt{n}}$, получаем

$$F_{Z_n}(x) \to \Phi(x), \ \forall x \in \mathbb{R},$$

где Φ(*x*) — функция распределения стандартного нормального распределения.

- Центральная предельная теорема в классической формулировке доказывается методом характеристических функций (теорема Леви о непрерывности).
- Вообще говоря, из сходимости функций распределения не вытекает сходимость плотностей. Тем не менее, в данном классическом случае имеет место.

Вопросы для самопроверки:

1. Сравните гипотезы многовариантности и предопределенности.
2. Сформулируйте принцип детерминизма Лапласа.
3. Опишите кратко историю развития теории вероятностей.
4. Что называется теорией вероятностей?
5. Что называется вероятностью?
6. Что называется пространством элементарных событий?
7. Определите понятие частоты событий.
8. Что такое функция распределения вероятностей?
9. Какое распределение называется нормальным?
10. Назовите параметры нормального распределения.
11. Что такое среднеквадратическое отклонение?
12. В чем заключается содержание закона больших чисел?
13. Сформулируйте усиленный закон больших чисел.
14. Приведите классическую формулировку центральной предельной теоремы.

Литература

1. Штейнгауз Г. Математический калейдоскоп/Пер. с польск. - М.: ГИТТЛ, 1949. - 143с.

2. Николис Г., Пригожин И. Самоорганизация в неравновесных структурах/Пер. с англ. - М.: Мир, 1979. - 327с.

3. Морозов Л.М., Петухов Г.Б., Сидоров В.Н. Методологические основы теории эффективности: Учебное пособие. - Л.: ВИКИ им. А.Ф.Можайского, 1982. - 236с.

4. Таранов П.С. Управление без тайн. - Донецк: Сталкер, 1997. - 448с.

5. Элементы теории испытаний и контроля технических систем/В.И. Городецкий, А.К. Дмитриев, В.М. Марков и др. Под ред. Р.М. Юсупова. - М.: Энергия, 1978. - 191с.

Приложение 1. Предельные теоремы теории вероятностей

Слабый закон больших чисел

Пусть есть бесконечная последовательность (последовательное перечисление) одинаково распределённых и некоррелированных случайных величин $\{X_i, \quad i=1,...,\infty\}$, определённых на одном вероятностном пространстве (Ω, F, P). То есть их ковариация $\operatorname{cov}(X_i, X_j) = 0, \ \forall i \neq j$. Пусть $\mathbb{E}X_i = \mu, \ \forall i \in \mathbb{N}$. Обозначим S_n выборочное среднее первых n членов:

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad n \in N .$$

Тогда $S_n \xrightarrow{P} \mu$.

Локальная Центральная предельная теорема

В предположениях классической формулировки, допустим в дополнение, что распределение случайных величин $\{X_i, \quad i=1,...,\infty\}$ абсолютно непрерывно, то есть оно имеет плотность. Тогда распределение Z_n также абсолютно непрерывно, и более того, $f_{Z_n}(x) \underset{n\to\infty}{\to} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, где $f_{Z_n}(x)$- плотность случайной величины Z_n, а в правой части стоит плотность стандартного нормального распределения.

Некоторые обобщения

Результат классической центральной предельной теоремы справедлив для ситуаций гораздо более общих, чем полная независимость и одинаковая распределённость.

Центральная предельная теорема Линденберга

Пусть независимые случайные величины $\{X_i, \quad i = 1,...,\infty\}$ определены на одном и том же вероятностном пространстве и имеют конечные математические ожидания и дисперсии: $\mathbb{E}[X_i] = \mu_i,\ \mathrm{D}[X_i] = \sigma_i^2$. Как и прежде построим частичные суммы

$S_n = \sum_{i=1}^{n} X_i$. Тогда в частности,

$$\mathbb{E}[S_n] = m_n = \sum_{i=1}^{n} \mu_i,\ \mathrm{D}[S_n] = s_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

Пусть выполняется условие Линденберга:

$$\forall \varepsilon > 0,\ \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left[\frac{(X_i-\mu_i)^2}{s_n^2} \mathbf{1}_{\{|X_i-\mu_i|>\varepsilon s_n\}}\right] = 0.$$

Тогда

$\frac{S_n - m_n}{s_n} \to N(0,1)$ по распределению при $n \to \infty$.

Центральная предельная теорема Ляпунова

Пусть выполнены базовые предположения Ц.П.Т. Линденберга. Пусть случайные величины $\{X_i\}$ имеют конечный третий момент. Тогда определена последовательность

$r_n^3 = \sum_{i=1}^{n} \mathbb{E}\left[|X_i - \mu_i|^3\right]$. Если предел $\lim_{n\to\infty} \frac{r_n}{s_n} = 0$ (условие Ляпунова), то

$\frac{S_n - m_n}{s_n} \to N(0,1)$ по распределению при $n \to \infty$.

Приложение 2. Учебники и монографии по теории вероятностей и математической статистике

Ахтямов, А. М. Теория вероятностей. — М.: Физматлит, 2009

Боровков, А. А. *Математическая статистика*, М.: Наука, 1984.

Боровков, А. А. *Теория вероятностей*, М.: Наука, 1986.

Булдык, Г. М. *Теория вероятностей и математическая статистика*, Мн., Высш. шк., 1989.

Булинский, А. В., Ширяев, А. Н. *Теория случайных процессов*, М.: Физматлит, 2003.

Бекарева, Н. Д. *Теория вероятностей. Конспект лекций*, Новосибирск НГТУ

Баврин, И. И. Высшая математика (Часть 2 «Элементы теории вероятностей и математической статистики»), М.: Наука, 2000.

Вентцель Е. С. *Теория вероятностей.* — М.: Наука, 1969. — 576 с.

Гмурман, В. Е. *Теория вероятностей и математическая статистика*: Учеб. пособие — 12-е изд., перераб.- М.: Высшее образование, 2006.-479 с.:ил (Основы наук).

Гмурман, В. Е. *Руководство к решению задач по теории вероятностей и математической статистике*: Учеб. пособие — 11-е изд., перераб. — М.: Высшее образование, 2006.-404 с. (Основы наук).

Гурский Е. И. *Сборник задач по теории вероятностей и математической статистике*, — Минск: Высшая школа, 1975.

П. Е. Данко, А. Г. Попов, Т. Я. Кожевников. Высшая математика в упражнениях и задачах. (В 2-х частях)- М.: Высш.шк, 1986.

Колемаев, В. А. и др. *Теория вероятностей и математическая статистика*, — М.: Высшая школа, 1991. http://www.iqlib.ru/book/preview/b0ce99dc4e 1741128564b81841 ae6ce0

Колмогоров, А. Н. *Основные понятия теории вероятностей*, М.: Наука, 1974.

Коршунов, Д. А., Фосс, С. Г. *Сборник задач и упражнений по теории вероятностей*, Новосибирск, 1997.

Коршунов, Д. А., Чернова, Н. И. *Сборник задач и упражнений по математической статистике*, Новосибирск. 2001.

Кремер Н. Ш. Теория вероятностей и математическая статистика: Учебник для ВУЗов. — 2- изд., перераб. и доп.-М:ЮНИТИ-ДАНА, 2004. — 573 с.

Кузнецов, А. В. *Применение критериев согласия при математическом моделировании экономических процессов*, Мн.: БГИНХ, 1991.

Лихолетов И. И., Мацкевич И. Е. *Руководство к решению задач по высшей математике, теории вероятностей и математической статистике*, Мн.: Выш. шк., 1976.

Лихолетов И. И. *Высшая математика, теория вероятностей и математическая статистика*, Мн.: Выш. шк., 1976.

Лоэв М.В *«Теория вероятностей»*, — М.: Издательство иностранной литературы, 1962.

Маньковский Б. Ю., Таблица вероятности

Мацкевич И. П., Свирид Г. П. *«Высшая математика. Теория вероятностей и математическая статистика»*, Мн.: Выш. шк., 1993.

Мацкевич И. П., Свирид Г. П., Булдык Г. М. *«Сборник задач и упражнений по высшей математике. Теория вероятностей и математическая статистика»*, Мн.: Выш. шк., 1996.

Мейер П.-А. Вероятность и потенциалы. Издательство Мир, Москва, 1973.

Млодинов Л. (Не)совершенная случайность

Прохоров, А. В., В. Г. Ушаков, Н. Г. Ушаков. *«Задачи по теории вероятностей»*, Наука. М.: 1986.

Прохоров Ю. В., Розанов Ю. А. *«Теория вероятностей»*, — М.: Наука, 1967.

Пугачев, В. С. *«Теория вероятностей и математическая статистика»*, Наука. М.: 1979.

Ротарь В. И., *«Теория вероятностей»*, — М.: Высшая школа, 1992.

Свешников А. А. и др., *Сборник задач по теории вероятностей, математической статистике и теории случайных функций*, — М.: Наука, 1970.

Свирид, Г. П., Макаренко, Я. С., Шевченко, Л. И. *Решение задач математической статистики на ПЭВМ*, Мн., Выш. шк., 1996.

Севастьянов Б. А., *Курс теории вероятностей и математической статистик»*, — М.: Наука, 1982.

Севастьянов, Б. А., Чистяков, В. П., Зубков, А. М. *Сборник задач по теории вероятностей»*, М.: Наука, 1986.

Соколенко А. И., *Высшая математика*, учебник. М.: Академия, 2002.

Феллер, В. *Введение в теорию вероятностей и её приложения*.

Хамитов, Г. П., Ведерникова, Т. И. *Вероятности и статистики*, БГУЭП. Иркутск.: 2006.

Чистяков, В. П. *Курс теории вероятностей*, М., 1982.

Шейнин О. Б. Теория вероятностей. Исторический очерк. Берлин: NG Ferlag, 2005, 329 с.

Ширяев, А. Н. *Вероятность*, Наука. М.: 1989.

ЛЕКЦИЯ 7

ПРОГНОСТИКА: СОВРЕМЕННЫЕ ТЕХНОЛОГИИ

1. Прогностика: взгляд в будущее

НАСТОЯЩЕГО не существует.

Настоящее – это бесконечно малый интервал времени, отторгаемый нашим сознанием от прошлого и будущего.

Настоящее не является реальностью, это некоторый виртуальный образ, формируемый игрой человеческого разума.

Видимое назначение разума – это обработка и анализ сенсорной информации с целью формирования управляющих решений на сознательном и бессознательном уровнях. Эффективность управляющих решений, определяющих не только благополучие, но и сам факт выживания homo sapiens в дарвинистском мире, определяется качеством анализа данных и формируемого на его основе прогноза.

При наличии качественного, достоверного прогноза развития ситуации, формирование эффективного решения не представляет труда. Достаточно лишь посмотреть на последствия реализации тех или иных управляющих решений и выбрать решение, в наибольшей степени отвечающее выбранному критерию качества.

Человек прогнозирует практически всегда. Все наши планы опираются на наши доморощенные прогнозы. Если прогноз верен, то планы реализуются

В настоящее время долгосрочные прогнозы бывают двух типов: плохие и очень плохие.

Что же мешает построить качественный прогноз? Прежде всего неопределенность, существующая в исходных данных относительно объекта исследования.

Ведь согласно принципу детерминизма Пьера-Симона Лапласа, динамика всех процессов определяется жесткими причинно-следственными связями типа «если – то».

Великий всезнайка, «демон Лапласа», знающий характер всех факторов влияния, смог бы спрогнозировать динамику развития любой системы на сколь угодно большой срок.

Значит ли это, что неопределенность, на позволяющая получить эффективный прогноз, существует лишь в силу ограниченности человеческих знаний о природе вещей и их взаимодействиях?

2. Классический прогноз

Рассмотрим две основных технологии классического прогнозирования: экспертные (они же эвристические) и математические, обычно реализуемые с помощью средств цифровой техники.

Насколько человек способен к строгому, количественному прогнозу. Пример на рис.1 иллюстрирует, что даже в полностью определенной ситуации с двумя простейшими факторами влияния, человеческий мозг не способен восстановить прогностическую динамику развития одномерного объекта.

Попытки использовать групповые *экспертные методы* прогнозирования также являются малоэффективными. Очевидно, что мнение лучшего эксперта будет заменена в этом случае менее качественной усредненной оценкой.

Математический аппарат традиционной прогностики включает в себя детерминированные и статистические модели. Для его реализации, теоретически, можно воспользоваться и готовыми программными пакетами, такими, как SPSS, Statistica, Statgraphics, Stadia, SAS, TimeLab и др. На практике готовые шаблонные формы редко позволяют с достаточной точностью формировать эффективные прогнозы.

Суть экстраполяционного прогноза хорошо известна и проиллюстрирована на рис. 2.

Наличие измерений позволяет перейти к математической модели динамики состояния. Прогнозирование сводится к статистической экстраполяции процесса эволюции состояния объекта прогнозирования. При этом, в идеальном случае, ошибка прогноза плавно растет с ростом горизонта прогноза.

Однако на практике, чаще всего, происходит то или иное скачкообразное изменения состояния объекта прогнозирования. Как правило, такой скачок происходит в силу неполноты мониторинга многообразных факторов влияния. Однако имеются и другие причины.

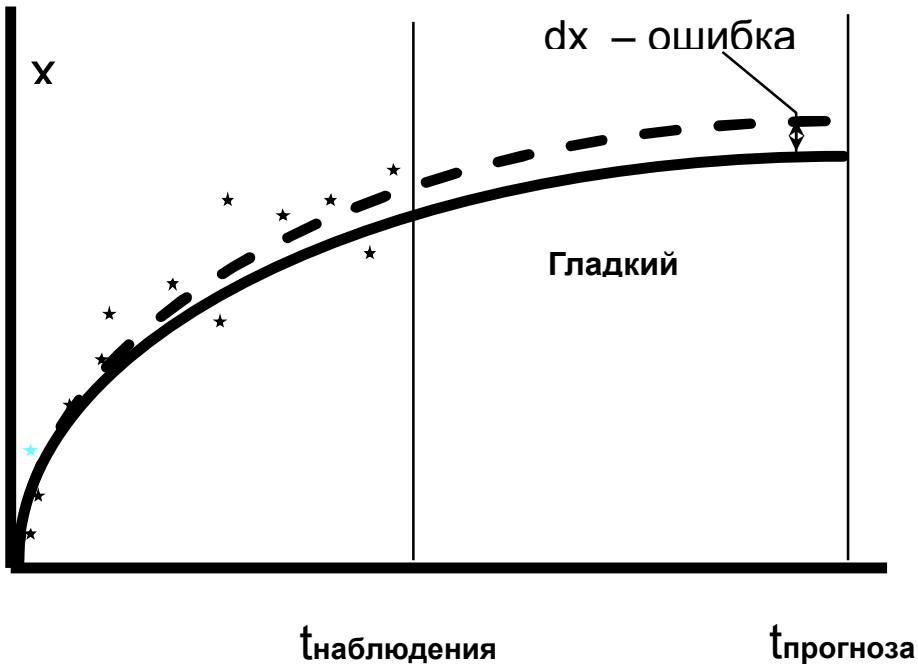


Рис. 2. Технология гладкого экстраполяционного прогноза

Проблемами для построения качественного, достоверного прогноза являются:

- неполнота и низкое качество «*сырых данных*» (Raw Data), полученных в процессе мониторинга и используемых для построения и коррекции математических моделей;
- малые выборки наблюдений;
- сверхбольшие объемы наблюдений;
- аномальные наблюдения;
- крайне высокие требования к точности идентификации протекающих процессов, связанные с соизмеримостью относительного выигрыша с флуктуационными характеристиками эволюции состояния объекта;
- необходимость значительных инвестиций, обусловленных сложностью стыковки и адаптации готовых комплексных систем прогнозирования с уже развернутыми средствами мониторинга;
- закрытость алгоритмического обеспечения, как коммерческого продукта, его недоступность для оперативной адаптации со стороны пользователей.

Специфика современных требований к системам прогнозирования состоит в сверхбольшой объем данных, разнородность и низкой структурированности данных, существенной глубине анализа, сложности интерпретируемость данных, доступности инструментария и др.

3. Прогностика: современные технологии

Проблемы формирования прогностической аналитики состоят в

- обнаружении идентификации скрытых факторов влияния;
- выявлении и идентификации скрытых взаимных связей;
- выявлении и идентификации тенденций изменения состояния,

т.е. совпадают с традиционными задачами DM.

На рис. 3 приведен вариант классификации задач, решаемых на основе данной методологии, а в табл. 1 – некоторые варианты программных продуктов, предназначенных для их решений

.

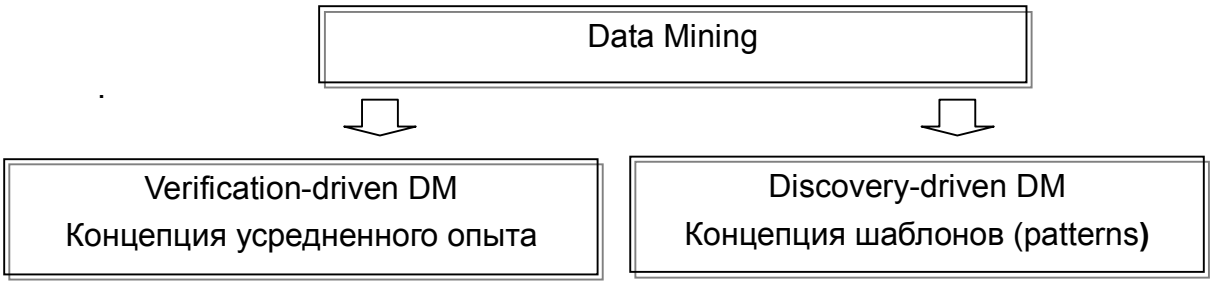


Рис. 3. Вариант классификации задач, решаемых на основе DM

Табл. 1. Варианты программных продуктов на основе DM

DM классы	Системы	Стоимость
Предметно-ориентированные аналитические системы	Скрининговые системы, ИС ЛПУ, ИС врача, ИС фельдшера, инф-справ. ИС и др.	$ 300-20000
Статистический анализ	SPSS, SAS, STATGRAPHICS, STATISTICA, STADIA	$1000-15000
Нейронные сети	BrainMarker, NeuroShell, OWL	$ 1500-8000
Ассоциации по аналогии	CBR, KATE Tools, Pattern Recognition Workbench	$1500 -10000
Деревья решений	See5/C5.0, Clementine, SIPINA, KnowledgeSEEKER	$1000 -10000
Эволюционное программирование	PolyAnalyst, NeuroShell	$1000 -5000
Генетические алгоритмы	GeneHunter	$1000
Алгоритмы ограниченного перебора	WizWhy	$4000
Системы визуализации многомерных данных	DataMiner3D	До $1000

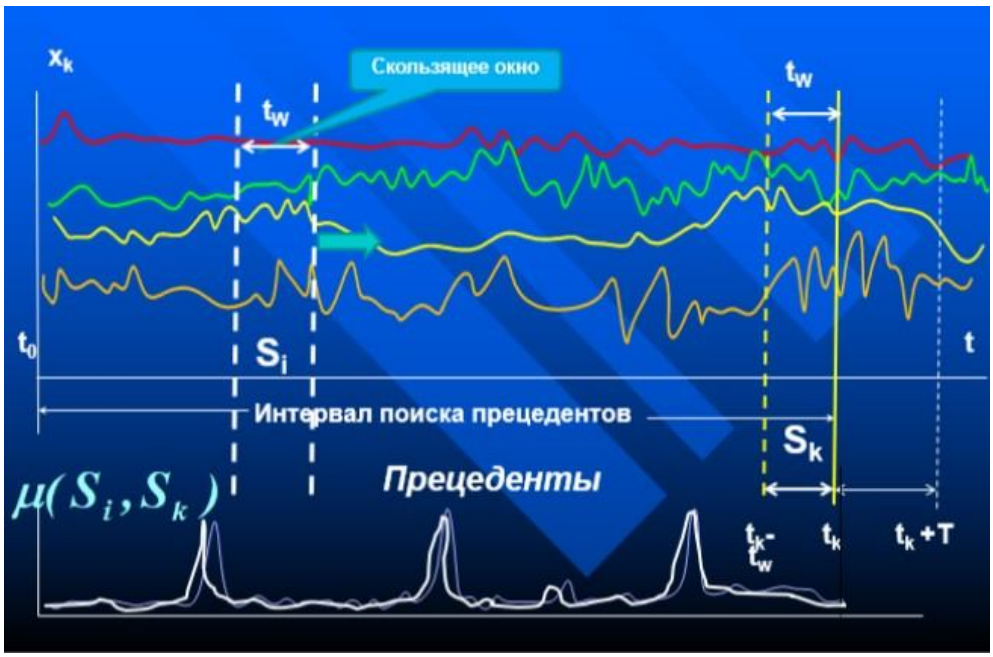


Рис. 4. Иллюстрация прецедентного анализа данных

Среди современных технологий прогноза, прежде всего, следует отметить технологию прецедентного анализа или метод динамических шаблонов – patterns. Суть метода проиллюстрирована на рис.4.

Частным случаем такого подхода является метод ближайшего соседа.

Конкретный пример реализации прецедентного анализа приведен на рис. 5.

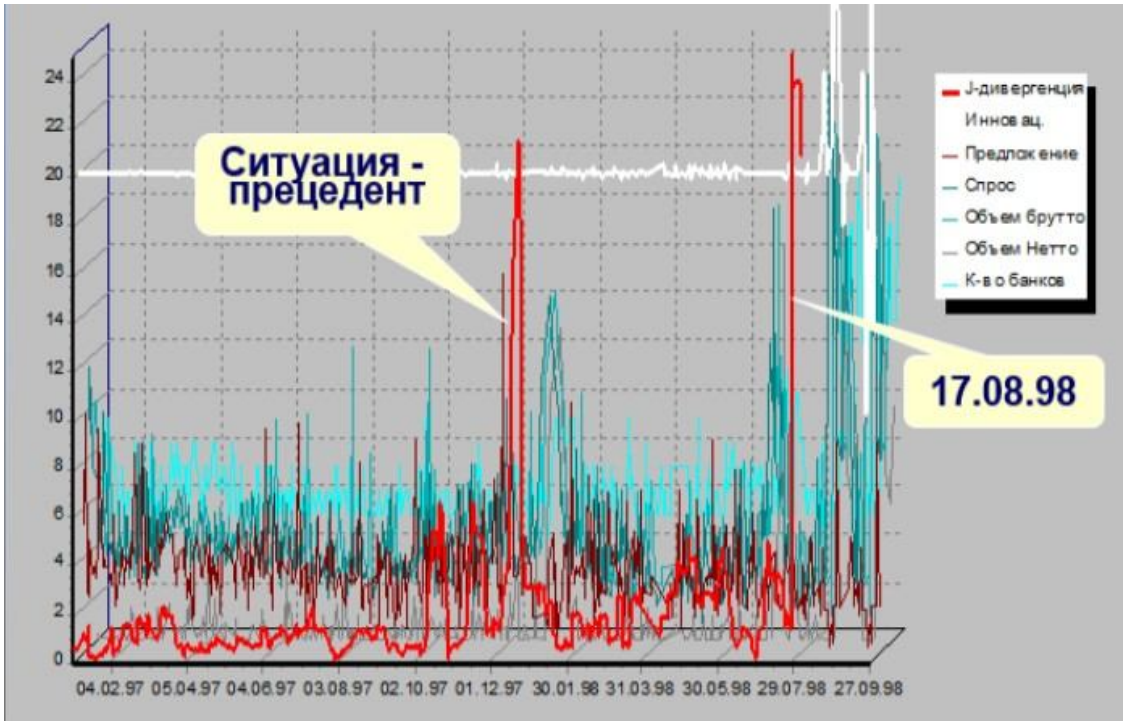


Рис. 5. Пример реализации прецедентного анализа

Особое значение в современной прогностике приобрели исследования в области динамики хаотических систем. Генезисом детерминированного хаоса яв-

ляется структурная неустойчивость открытых нелинейных систем в так называемых точках бифуркаций. Соответствующая иллюстрация приведена на рис. 6.

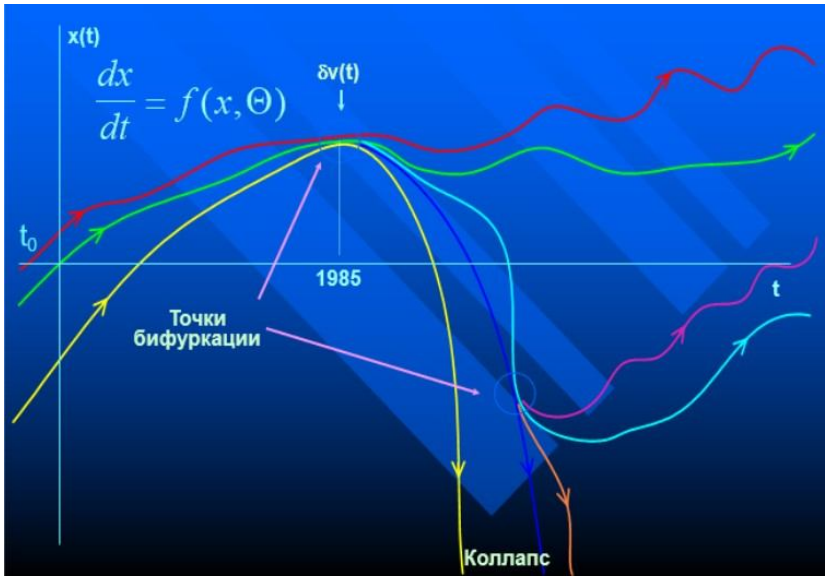


Рис. 6. Пример структурная неустойчивость открытых нелинейных систем

Третьим перспективным направлением анализа данных и прогнозирования являются искусственные нейронные сети. Модель нервной клетки нейрона приведена на рис.7**,** простейшая модель нейронной сети, двухслойный персептрон – на рис.8.

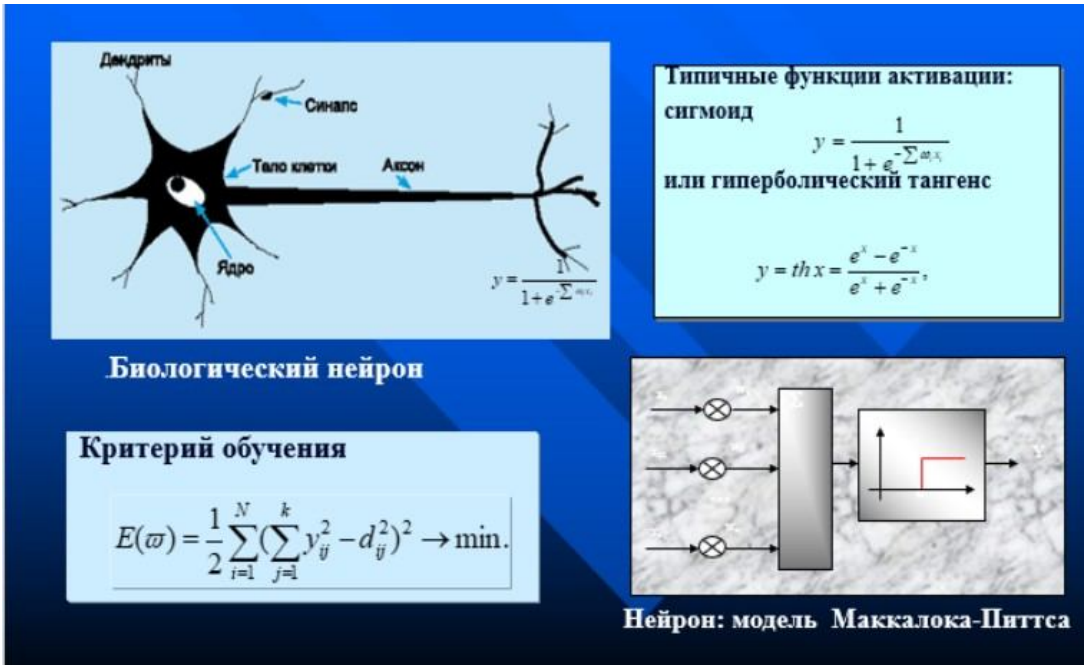


Рис. 7. Модель нейрона

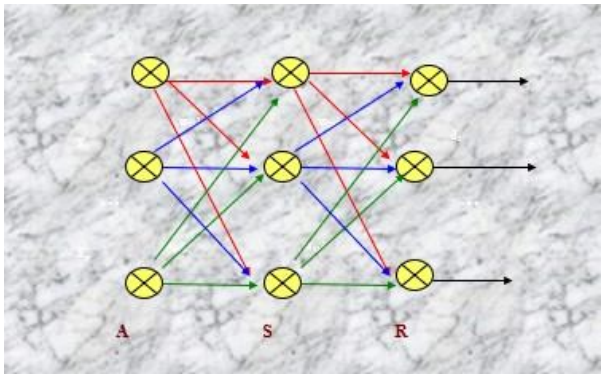


Рис. 8. Персептрон Розенблатта

Четвертое направление развития когнитивной прогностики основано на технологии эволюционного моделирования. Идея эволюционного подхода основана на дарвинистской теории селекции и отбора и представлена на рис. 9.

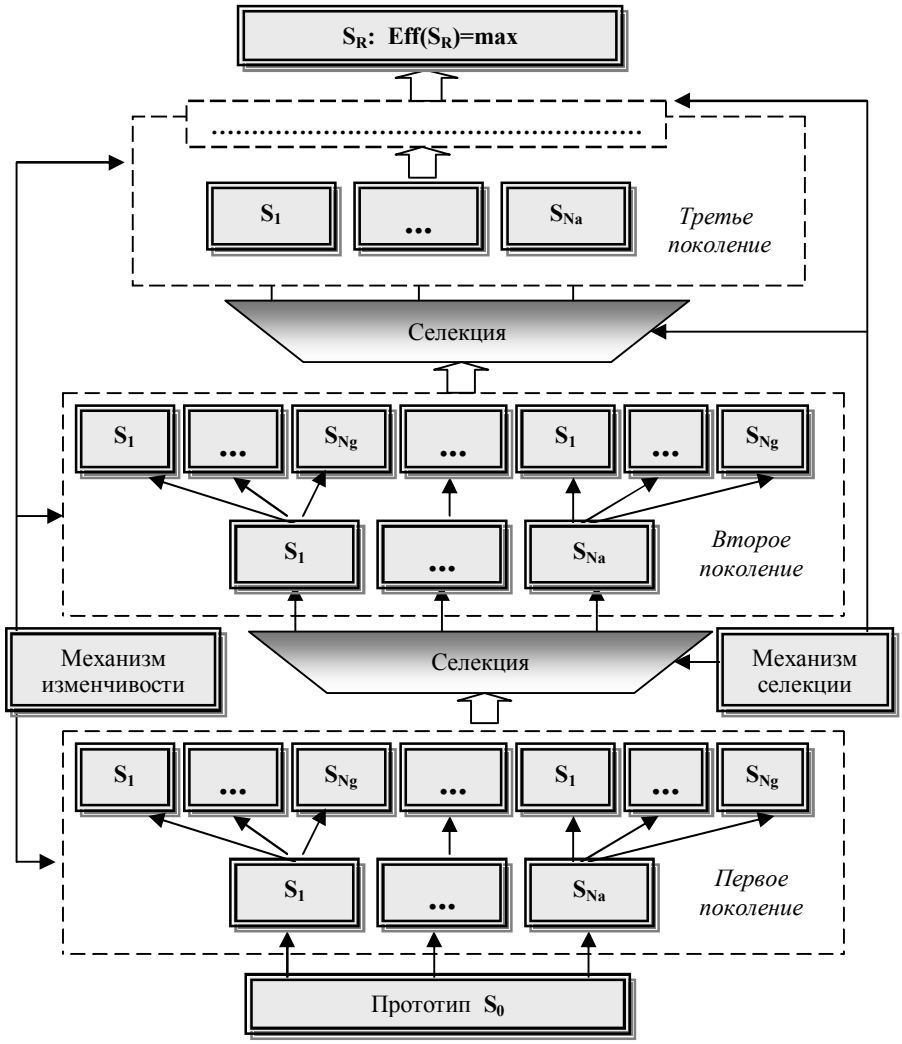


Рис. 9. Общая функциональная структура алгоритма эволюционной оптимизации

Некоторым развитием эволюционного моделирования являются генетические алгоритмы, представленные нас слайде 10.

Рис. 10. Генетические алгоритмы

Вопросы для самопроверки:

1. Сформулируйте принцип детерминизма Лапласа.
2. Что называется прогнозом? Чем прогноз отличается от предсказания?
3. Назовите две основных технологии классического прогнозирования.
4. В чем состоит экстраполяционный прогноз? Экспертный прогноз?
5. Перечислите проблемы построения качественного, достоверного прогноза.
6. Приведите вариант классификации задач, решаемых на основе данной Data Mining.
7. Приведите примеры программных продуктов на основе DM.
8. Объясните технологию прецедентного анализа.
9. Назовите генезис детерминированного хаоса.
10. Опишите модель нейрона.
11. Приведите пример искусственной нейронной сети.
12. Опишите технологию эволюционного моделирования.
13. Опишите технологию генетических алгоритмов.

II. СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

ЛЕКЦИЯ 8

ОСНОВЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

1. Основные понятия статистического анализа данных

Определение 1. Математическая статистика - наука, выявляющая закономерности повторяющихся случайных явлений на основе обработки статистических данных, полученных в результате наблюдений.

Определение 2. Математическая статистика - наука, разрабатывающая математические методы систематизации и использования статистических данных для научных и практических выводов.

Будем различать три основных блока функциональностей, относящихся к математической статистике:

- Дескриптивная статистика - совокупность эмпирических методов, используемых для визуализации и интерпретации данных (расчет выборочных характеристик, таблицы, диаграммы, графики и т. д.);
- Анализ и установлением связей и закономерностей, которым подчинены повторяющиеся случайные явления, на основе обработки статистических данных, полученных в результате наблюдений;
- Классификация и распознавание образов.

Терминальным пользователем статистических методов обычно являются системы прогнозирования и системы поддержки и формирования управляющих решений.

Основными задачами математической статистики являются:

1. Разработка методов анализа данных в зависимости от целей исследования, к которым относятся:
   - оценка неизвестной вероятности события, неизвестной функции распределения и ее параметров;
   - оценка зависимостей от случайных величин и т.д.;
   - проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения;
2. Синтез алгоритмов прикладной статистики для решения задач выявления взаимосвязей, трендов, прогнозирования, поддержки принятия решений и т.п.

Для решения этих задач необходимо выбрать из большой совокупности однородных объектов ограниченное количество объектов, по результатам изучения которых можно сделать прогноз относительно исследуемого признака этих объектов.

Введем *основные понятия* математической статистики.

Генеральная совокупность – все множество имеющихся данных, наблюдений или объектов, относящихся к изучаемому явлению.

Выборка – набор наблюдений или объектов, случайно отобранных из генеральной совокупности.

Объем генеральной совокупности N и объем выборки n – число наблюдений или объектов в рассматриваемых совокупностях.

Виды выборки:

Повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность;

Бесповторная – отобранный объект в генеральную совокупность не возвращается.

Замечание. Для того, чтобы по исследованию выборки можно было сделать выводы о поведении интересующего нас признака генеральной совокупности, нужно, чтобы выборка правильно представляла изучаемые свойства генеральной совокупности, то есть была *репрезентативной* (представительной).

Для выполнения этого условия, в частности, необходимо, чтобы, учитывая закон больших чисел, каждый объект был выбран случайно, причем для любого объекта вероятность попасть в выборку одинакова.

Первичная обработка результатов. Пусть интересующая нас дискретная случайная величина X принимает в выборке значение x_1 m_1 раз, x_2 – m_2 раз, …, x_k – m_k раз, причем $\sum_{i=1}^{k} m_k = n$, где n – объем выборки. Тогда наблюдаемые значения случайной величины x_1, x_2,…, x_k называют наблюдениями или вариантами, а m_1, m_2,…, m_k – частотами.

Если разделить каждую частоту на объем выборки n, то получим *относительные частоты* $w_i = \frac{m_i}{n}$.

Определение. Перечень наблюдений и соответствующих им частот или относительных частот называют *статистическим рядом:*

x_i	x_1	x_2	…	x_k
n_i	m_1	m_2	…	m_k
w_i	w_1	w_2	…	w_k

Пример. При проведении 20 бросков игральной кости число выпадений очков оказалось равным 2, 2, 5, 1, 2, 3, 2, 3, 3, 1, 5, 4, 4, 2, 1, 3, 2, 3, 6, 4. Статистический ряд для абсолютных и относительных частот имеет вид:

x_i	1	2	3	4	5	6
m_i	3	6	5	3	2	1
w_i	0,15	0,3	0,25	0,15	0,1	0,05

Определение. Последовательность наблюдений, записанных в порядке возрастания или убывания $x_{(1)}$, $x_{(2)}$,…, $x_{(к)}$: $x_{(1)} <= x_{(2)} <= \ldots <= x_{(к)}$ или убывания $x_{(1)}$, $x_{(2)}$,…, $x_{(к)}$: $x_{(1)} >= x_{(2)} >= \ldots >= x_{(к)}$ называют *вариационным рядом.*

Пример. Используем данные предыдущего примера: 1, 1, 4, 0, 1, 2, 1, 2, 2, 0, 5, 3, 3, 1, 0, 2, 2, 3, 4, 1. Соответствующий вариационный ряд имеет вид: 0, 1, 2, 3, 4, 5.

Определение. Наблюдения, образующие вариационный ряд

$$x_{(1)},\ x_{(2)},\ \ldots,\ x_{(k)}$$

называются *порядковыми статистиками*.

Определение. Номера порядковых статистик в вариационном ряду называются их *рангами*.

2. Группированные данные

В случае, когда значения признака являются *непрерывными,* удобно использовать группированную выборку.

Для ее получения интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько равных частичных интервалов длиной h, а затем находят для каждого частичного интервала n_i – сумму частот наблюдений, попавших в i-й интервал. Составленная по этим результатам таблица называется группированным статистическим рядом:

Номера интервалов	1	2	…	k
Границы интервалов	$(a,\ a+h)$	$(a+h,\ a+2h)$	…	$(b-h,\ b)$
Сумма частот наблюдений, попавших в интервал	m_1	m_2	…	m_k

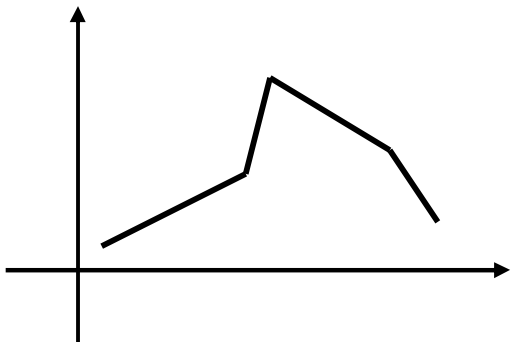


Рис. 1. Полигон частот

Для наглядного представления о поведении исследуемой случайной величины в выборке можно строить различные графики.

Один из них – полигон частот: ломаная, отрезки которой соединяют точки с координатами $(x_1, m_1), (x_2, m_2), \ldots, (x_k, m_k)$, где x_i *откладываются* на оси абсцисс, а m_i – на оси ординат.

Если на оси ординат откладывать не абсолютные (m_i), а относительные (w_i) частоты, то получим полигон относительных частот (рис.1).

3. Выборочная функция распределения и гистограмма

По аналогии с функцией распределения случайной величины можно задать относительную частоту события X ***< x.***

Определение*. Выборочной (эмпирической) функцией распределения* называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события X <x. Таким образом,

$$F^*(x) = \frac{m_x}{n},$$

где m_x – число наблюдений, меньших x, n – объем выборки.

Замечание. В отличие от эмпирической функции распределения, найденной опытным путем, функцию распределения $F(x)$ генеральной совокупности называют *теоретической функцией распределения*.

$F(x)$ определяет вероятность события $X<x$, а $F^*(x)$ – его относительную частоту. При достаточно больших n, как следует из теоремы Бернулли, $F^*(x)$ стремится по вероятности к $F(x)$.

Из определения эмпирической функции распределения видно, что ее свойства совпадают со свойствами $F(x)$, а именно:

1. $0 \le F^*(x) \le 1$.
2. $F^*(x)$ – неубывающая функция.
3. Если x_1 – наименьшее наблюдение, то $F^*(x) = 0$ при $x \le x_1$; если $x_к$ – наибольшее наблюдение, то $F^*(x) = 1$ при $x > x_к$.

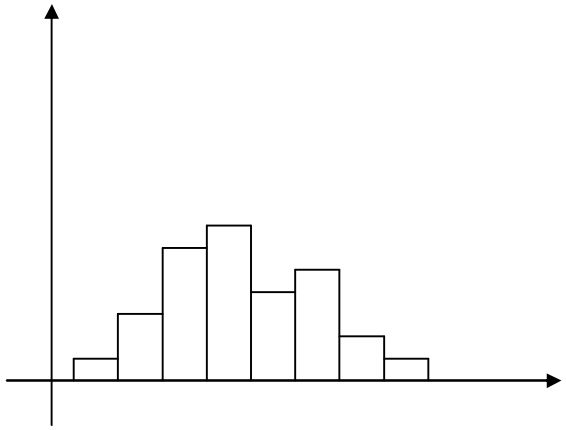


Рис.2.

Для *непрерывного* признака графической иллюстрацией служит гистограмма, то есть ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат локальные интервалы длиной h, а высотами – отрезки длиной n_i / h (гистограмма частот) или w_i/h (гистограмма относительных частот). В первом случае площадь гистограммы равна объему выборки, во втором – единице (рис.2).

4. Оценки параметра положения: в*ыборочное среднее, оценки моды и медианы*

Одна из задач математической статистики: по имеющейся выборке оценить значения числовых характеристик исследуемой случайной величины.

Определение. Выборочным средним называется среднее арифметическое значений случайной величины, принимаемых в выборке:

$$\overline{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n}, \qquad (1)$$

где x_i – наблюдения, n_i - частоты.

Замечание. Выборочное среднее служит для оценки математического ожидания исследуемой случайной величины. В дальнейшем будет рассмотрен вопрос, насколько точной является такая оценка.

Другими характеристиками статистического ряда являются:

- мода Mod – наблюдение, имеющая наибольшую частоту: $\mathrm{Mod} = x_{k^*} : \quad m_{k^*} = \max(p_1, \ldots, p_n)$;

- медиана Med - наблюдение, которая делит вариационный ряд на две части, равные по числу наблюдений: $\mathrm{Med}=\begin{cases} x_{k+1}, & n=2k+1; \\ \dfrac{x_k+x_{k+1}}{2}, & n=2k \end{cases}$,

т.е., если число наблюдений нечетно (n=2k+1), то Med=x_{k+1}, а при четном n=2k $\boldsymbol{Med=\dfrac{x_k+x_{k+1}}{2}}$.

5. Оценки параметра масштаба: оценки *дисперсии, начальных и центральных моментов*

Определение. Выборочной дисперсией называется

$$\hat{D}_n=\frac{\sum_{i=1}^{n}(x_i-\overline{x}_n)^2}{n}=\frac{\sum_{i=1}^{k}m_i(x_i-\overline{x}_n)^2}{n}, \qquad (2)$$

Выборочным средним квадратическим отклонением –

$$\hat{\sigma}_n=\sqrt{\hat{D}_n}. \qquad (3)$$

Так же, как в теории случайных величин, можно доказать, что справедлива следующая формула для вычисления выборочной дисперсии:

$$\hat{D}=\overline{x^2}-(\overline{x})^2. \qquad (4)$$

Пример. Найдем числовые характеристики выборки, заданной статистическим рядом

x_i	2	5	7	8
m_i	3	8	7	2

$$\overline{x}_n=\frac{2\cdot3+5\cdot8+7\cdot7+8\cdot2}{20}=5{,}55;$$

$$\hat{D}_n=\frac{4\cdot3+25\cdot8+49\cdot7+64\cdot2}{20}-5{,}55^2=3{,}3475;$$

$$\hat{\sigma}_n=\sqrt{3{,}3475}=1{,}83.$$

$$\mathrm{Mo}=5;\quad \mathrm{Me}=\frac{5+7}{2}=6.$$

Оценки начальных и центральных моментов (так называемые эмпирические моменты) определяются аналогично соответствующим теоретическим моментам:

- начальным эмпирическим моментом порядка k называется

$$\boldsymbol{\hat{\alpha}_n^k=\frac{1}{n}\sum_{i=1}^{n}x_i^k} \quad . \qquad (5)$$

При наличии повторяющихся значений эту формулу можно записать в виде

$$\hat{\alpha}_n^k = \frac{\sum m_i x_i^k}{n}$$

В частности, $\hat{\alpha}_1 = \frac{\sum n_i x_i}{n} = \bar{x}_n$, то есть начальный эмпирический момент первого порядка равен выборочному среднему.

- центральным эмпирическим моментом порядка k называется

$$\hat{\mu}_n^k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^k \quad \text{или} \quad \hat{\mu}_n^k = \frac{\sum m_i (x_i - \bar{x}_n)^k}{n} \qquad (6)$$

В частности, $\hat{\mu}_2 = \frac{\sum n_i (x_i - \bar{x}_B)^2}{n} = D_n$, то есть центральный эмпирический момент второго порядка равен выборочной дисперсии.

6. Свойства оценок

Получив статистические оценки параметров распределения (выборочное среднее, выборочную дисперсию и т.д.), нужно убедиться, что они в достаточной степени служат приближением соответствующих характеристик генеральной совокупности. Определим требования, которые должны при этом выполняться.

Пусть $\hat{\theta}$ - статистическая оценка неизвестного параметра θ теоретического распределения. Извлечем из генеральной совокупности *k* выборок одного и того же объема *n* и вычислим для каждой из них оценку параметра θ: $\hat{\Theta}_1, \hat{\Theta}_2, ..., \hat{\Theta}_n$. Тогда оценку $\hat{\theta}$ можно рассматривать как случайную величину, принимающую возможные значения $\hat{\Theta}_1, \hat{\Theta}_2, ..., \hat{\Theta}_n$. Если математическое ожидание Θ* не равно оцениваемому параметру, мы будем получать при вычислении оценок систематические ошибки одного знака (с избытком, если $E(\hat{\theta}) > \theta$, и с недостатком, если $E(\hat{\theta}) < \theta$. Следовательно, необходимым условием отсутствия систематических ошибок является требование $E(\hat{\theta}) = \Theta$.

Определение. Статистическая оценка $\hat{\theta}$ называется *несмещенно*й, если ее математическое ожидание равно оцениваемому параметру θ при любом объеме выборки: $E(\hat{\theta}) = \theta$.

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Несмещенность не является достаточным условием хорошего приближения к истинному значению оцениваемого параметра. Если при этом возможные значения $\hat{\theta}$ могут значительно отклоняться от среднего значения, то есть дисперсия $\hat{\theta}$ велика, то значение, найденное по данным одной выборки, может значительно отличаться от оцениваемого параметра. Следовательно, требуется наложить ограничения на дисперсию.

Определение. Статистическая оценка называется эффективной, если она при заданном объеме выборки ***n*** имеет наименьшую возможную дисперсию

$D(\hat{\theta}) = min$.

При рассмотрении выборок большого объема к статистическим оценкам предъявляется еще и требование состоятельности.

***Определение**. Состоятельной* называется статистическая оценка, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру: $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$

Заметим, что если оценка несмещенная, то она будет состоятельной, если при $n \to \infty$ ее дисперсия стремится к ***0***.

Убедимся, что $\bar{x}_n$ представляет собой несмещенную оценку математического ожидания $E(x)$.

Будем рассматривать $\bar{x}_n$ как случайную величину, а $x_1, x_2, ..., x_n$, то есть значения исследуемой случайной величины, составляющие выборку, – как реализации независимых, одинаково распределенных случайные величин $X_1, X_2, ..., X_n$, имеющих одинаковое математическое ожидание ***a***. Из свойств математического ожидания следует, что

$$E(\bar{X}_n) = E\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{na}{n} = a.$$

Но, поскольку каждая из величин X_1, X_2,…, X_n имеет такое же распределение, что и генеральная совокупность, $a = E(X)$, то есть $E(\bar{X}_n) = E(X)$, что и требовалось доказать.

Выборочное среднее является не только несмещенной, но и состоятельной оценкой математического ожидания. Если предположить, что $X_1, X_2, ..., X_n$ имеют ограниченные дисперсии, то из теоремы Чебышева следует, что их среднее арифметическое, то есть $\bar{X}_n$, при увеличении ***n*** стремится по вероятности к математическому ожиданию каждой их величин, то есть к $E(x)$. Следовательно, выборочное среднее есть состоятельная оценка математического ожидания.

В отличие от выборочного среднего, выборочная дисперсия является ***смещенной*** оценкой дисперсии генеральной совокупности. Можно доказать, что

$$E(\hat{D}_n) = \frac{n-1}{n} D \ ,$$

где ***D*** – истинное значение дисперсии генеральной совокупности.

Можно предложить другую оценку дисперсии – *исправленную дисперсию* s^2, вычисляемую по формуле

$$s^2 = \frac{n}{n-1}\hat{D}_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n-1} \quad \text{или} \quad s^2 = \frac{n}{n-1}\hat{D}_n = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}{n-1}.$$

Такая оценка будет являться несмещенной. Ей соответствует *исправленное среднее квадратическое отклонение*

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n-1}} \text{ или } s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}{n-1}}.$$

Вопросы для самопроверки:

1. Что такое математическая статистика?
2. Что изучает математическая статистика?
3. Основные задачи математической статистики?
4. Что называется генеральной совокупностью? Выборкой?
5. Какие бывают виды выборок?
6. Какая выборка является репрезентативной?
7. Что называется наблюдением? Статистическим рядом? Вариационным рядом? Рангом?
8. Как осуществляется группировка данных?
9. Что называется группированным статистическим рядом? Полигоном частот? Полигоном относительных частот?
10. Дайте определение выборочной (эмпирической) функцией распределения.
11. Напишите соотношения для оценок положения: в*ыборочного среднего, оценки моды и медианы.*
12. Напишите соотношения для оценок параметра масштаба: оценки дисперсии, начальных и центральных моментов.
13. Перечислите основные свойства статистических оценок.

ЛЕКЦИЯ 9.

МЕТОД МОНТЕ-КАРЛО (МЕТОД СТАТИСТИЧЕСКИХ ИСПЫТАНИЙ)

1. Постановка задачи

Метод Монте-Карло получил свое название в честь европейской столицы азартных игр, в которых случайность является определяющим фактором в получении выигрыша.

Типовой задачей, решаемой методом Монте-Карло, является поиск значения неизвестной неслучайной величины, *a* на основе розыгрыша (генерации) значений случайной величины *X*, числовые характеристики распределения которой связаны с искомой величиной.

Например, формируются реализации случайная величина *X*, математическое ожидание которой E_X равно *a.* В этом случае для выборки из *n* значений *X*, полученных в *n* испытаниях, вычисляется выборочное среднее:

$$\overline{x} = \frac{\sum x_i}{n},$$

которое принимается в качестве оценки искомого числа *a:* $a \approx \hat{a} = \overline{x}.$

Этот метод требует проведения большого числа испытаний, поэтому его иначе называют *методом статистических испытаний.*

Теория метода Монте-Карло исследует такие вопросы, как:

- как наиболее целесообразно выбрать случайную величину *X*,
- как найти ее возможные значения,
- как уменьшить дисперсию используемых случайных величин, чтобы погрешность при замене a на $\hat{a}$ была возможно меньшей.

Генерация возможных значений *X* называют *разыгрыванием случайной величины*. Рассмотрим некоторые способы разыгрывания случайных величин и выясним, как оценить допускаемую при этом ошибку.

2. Разыгрывание дискретной случайных величин

Определение. *Случайными числами* называют возможные значения ***r*** непрерывной случайной величины ***R***, распределенной равномерно в интервале **(0; 1).**

2.1. Общая схема разыгрывания дискретной случайной величины

Пусть требуется разыграть дискретную случайную величину *X*, то есть получить последовательность ее возможных значений, зная закон распределения *X*:

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n .

Рассмотрим равномерно распределенную в (0, 1) случайную величину *R*:

$R \in U[0, 1]$

и разобьем интервал (0, 1) точками с координатами p_1, p_1+p_2, …, $p_1+p_2+\ldots+p_{n-1}$ на n частичных интервалов $\Delta_1,\Delta_2,...,\Delta_n$, длины которых равны вероятностям с теми же индексами.

Теорема. Если каждому случайному числу $r_j \in U[0,1]$, $0 \le r_j < 1$, которое попало в интервал Δ_i, ставить в соответствие возможное значение x_j $r_j \leftrightarrow x_j$, $\forall j = 1,...,n$, то разыгрываемая величина будет иметь заданный закон распределения:

X	x_1	x_2 …	x_n
p	p_1	p_2 …	p_n.

Доказательство.

Возможные значения полученной случайной величины совпадают с множеством x_1, x_2 ,… x_n, так как число интервалов равно n, а при попадании r_j в интервал Δ_i случайная величина может принимать только одно из значений x_1, x_2 ,… x_n.

Так как *R* распределена равномерно, то вероятность ее попадания в каждый интервал равна его длине, откуда следует, что каждому значению x_i соответствует вероятность p_i. Таким образом, разыгрываемая случайная величина имеет заданный закон распределения.

Пример. Разыграть *10* значений дискретной случайной величины *X*, закон распределения которой имеет вид:

X	2	3	6	8
p	0,1	0,3	0,5	0,1

Решение. Разобьем интервал (0, 1) на частичные интервалы: Δ_1- [0; 0,1), Δ_2 – [0,1; 0,4), Δ_3 - [0,4; 0,9), Δ_4 – [0,9; 1].

Выпишем из таблицы случайных чисел *10* чисел: *0,09; 0,73; 0,25; 0,33; 0,76; 0,52; 0,01; 0,35; 0,86; 0,34*. Первое и седьмое числа лежат на интервале Δ_1, следовательно, в этих случаях разыгрываемая случайная величина приняла значение $x_1 = 2$; третье, четвертое, восьмое и десятое числа попали в интервал Δ_2, что соответствует $x_2 = 3$; второе, пятое, шестое и девятое числа оказались в интервале Δ_3 – при этом $X = x_3 = 6$; на последний интервал не попало ни одного числа.

Итак, разыгранные возможные значения *X* таковы: *2, 6, 3, 3, 6, 6, 2, 3, 6, 3.*

2.2. Разыгрывание противоположных событий

Пусть требуется разыграть испытания, в каждом из которых событие *A* появляется с известной вероятностью *p*.

Рассмотрим дискретную случайную величину *X*, принимающую значения 1 (в случае, если событие *A* произошло) с вероятностью *p* и 0 (если *A* не произошло) с вероятностью $q=1-p$. Затем разыграем эту случайную величину так, как было предложено в предыдущем пункте.

Пример. Разыграть *10* испытаний, в каждом из которых событие *A* появляется с вероятностью *0,3.*

Решение. Для случайной величины *X* с законом распределения

X	1	0
p	0,3	0,7

получим интервалы Δ_1 – (0; 0,3) и Δ_2 – (0,3; 1). Используем ту же выборку случайных чисел, что и в предыдущем примере, для которой в интервал Δ_1 попадают числа №№1,3 и 7, а остальные – в интервал Δ_2. Следовательно, можно считать, что событие *A* произошло в первом, третьем и седьмом испытаниях, а в остальных – не произошло.

2.3. Разыгрывание полной группы событий

Если события A_1, A_2, …, A_n, вероятности которых равны p_1 , p_2 ,… p_n, образуют полную группу, то для из разыгрывания (то есть моделирования последовательности их появлений в серии испытаний) можно разыграть дискретную случайную величину *X* с законом распределения *X:* 1 2 … *n,* сделав это так же, как в пункте 1. При этом считаем, что *p:* p_1 p_2 … p_n

если *X* принимает значение $x_i = i$, то в данном испытании произошло событие A_i.

3. Разыгрывание непрерывной случайной величины

3.1. Метод обратных функций.

Пусть требуется разыграть непрерывную случайную величину *X*, то есть получить последовательность ее возможных значений x_i (i = 1, 2, …, n), зная функцию распределения $F(x)$.

Теорема. Если r_i – случайное число, то возможное значение x_i разыгрываемой непрерывной случайной величины *X* с заданной функцией распределения $F(x)$, соответствующее r_i , является корнем уравнения

$$F(x_i) = r_i \qquad (1)$$

Доказательство.

1. Так как $F(x)$ монотонно возрастает в интервале от *0* до *1*, то найдется (причем единственное) значение аргумента x_i , при котором функция распределения примет значение r_i . Значит, уравнение (1) имеет единственное решение $x_i = F^{-1}(r_i)$, где F^{-1} - функция, обратная к *F :*

$F(x)$ монотонна на [0, 1] $\Rightarrow$ $\exists!\, x_i :\ x_i = F^{-1}(r_i)$.

.

2. Докажем, что $P\{X \in (c, d)\} = F(d) - F(c)$

В силу монотонности **F(x)** и того, что $F(x_i) = r_i$,

$c < x_i < d \Leftrightarrow F(c) < r_i < F(d)$

следовательно, $P(c < X < d) = P(F(c) < R < F(d)) = F(d) - F(c).$

Пример.

Разыграть ***3*** возможных значения непрерывной случайной величины ***X,*** распределенной равномерно в интервале ***(5; 8)***.

Решение.

$F(X) = \dfrac{x-5}{3}$, то есть требуется решить уравнение $\dfrac{x_i - 5}{3} = r_i,\ x_i = 3r_i + 5.$

Выберем *3* случайных числа: r= *0,23; 0,09; 0,56* и подставим их в это уравнение. Получим соответствующие возможные значения *X:* $x_1 = 5{,}69;\ x_2 = 5{,}27;\ x_3 = 6{,}68.$

3.2. Метод суперпозиции

Если функция распределения разыгрываемой случайной величины может быть представлена в виде линейной комбинации двух функций распределения:

$$F(x) = C_1 F_1(x) + C_2 F_2(x) \qquad (C_{1,2} > 0), \qquad (2)$$

то $C_1 + C_2 = 1$, так как при $x \to \infty$ $F(x) \to 1$.

Введем вспомогательную дискретную случайную величину *Z* с законом распределения *Z:* 1 2. Выберем *2* независимых случайных числа r_1 и r_2 и разыграем возможное

p: C_1 C_2;

значение *Z* по числу r_1 (см. пункт 1). Если *Z*=1, то ищем искомое возможное значение *X* из уравнения $F_1(x) = r_1$, а если *Z*=2, то решаем уравнение $F_2(x) = r_2$.

Можно доказать, что при этом функция распределения разыгрываемой случайной величины равна заданной функции распределения.

Пример: засоренное нормальное распределение (модель Хьюбера):

$F(x) = (1-\varepsilon)N(0,1) + \varepsilon F(0, \sigma^2) \qquad \varepsilon \in [0,1], \quad \sigma^2 > 1.$

3.3. Приближенное разыгрывание нормальной случайной величины

Так как для *R*, равномерно распределенной в (0, 1),

$$E(R) = \frac{1}{2}, \quad D(R) = \frac{1}{12},$$

то для суммы *n* независимых, равномерно распределенных в интервале (0, 1) случайных величин $\sum_{j=1}^{n} R_j$ $E\left(\sum_{j=1}^{n} R_j\right) = \frac{n}{2}$, $D\left(\sum_{j=1}^{n} R_j\right) = \frac{n}{12}$, $\sigma = \sqrt{\frac{n}{12}}$.

Тогда в силу центральной предельной теоремы нормированная случайная величина $\dfrac{\sum_{j=1}^{n} R_j - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$ при $n \to \infty$ будет иметь распределение, близкое к нормальному с параметрами *a*=0 и σ =1.

В частности, достаточно хорошее приближение получается при n = 12: $\sum_{j=1}^{12} R_j - 6.$

Итак, чтобы разыграть возможное значение нормированной нормальной случайной величины *x*, надо сложить 12 независимых случайных чисел и из суммы вычесть 6.

4. Оценка погрешности метода Монте-Карло

Если поставить задачу определения верхней границы допускаемой ошибки δ с заданной доверительной вероятностью γ, то есть поиска числа δ, для которого

$$P(|\overline{X} - a| \le \delta) = \gamma ,$$

то получим известную задачу определения *доверительного интервала* для математического ожидания генеральной совокупности. Воспользуемся результатами решения этой задачи для следующих случаев:

1) случайная величины *X* распределена нормально и известно ее среднее квадратическое отклонение: $X \in N(a, \sigma^2)$. Тогда из формулы для доверительного оценивания, получаем: $\delta = \frac{t\sigma}{\sqrt{n}}$, где *n* – число испытаний, σ - известное среднее квадратическое отклонение, а *t* – аргумент функции Лапласа, при котором $\Phi(t) = \gamma/2$.

2) случайная величина *X* распределена нормально с неизвестным σ. Воспользуемся формулой для интервального оценивания $\delta = \frac{t_\gamma s}{\sqrt{n}}$, где *s* – исправленное выборочное среднее квадратическое отклонение, а t_γ определяется по соответствующей таблице распределения Стьюдента $S(k, \gamma)$.

3) Если случайная величина распределена по иному закону $F \neq N(a, \sigma^2)$, то при достаточно большом количестве испытаний ($n > 30$) можно использовать для оценки δ предыдущие формулы, так как при $n \to \infty$ распределение Стьюдента стремится к нормальному $S(k, \gamma) \xrightarrow[n \to \infty]{} N$, и границы интервалов, полученные по ранее приведенным формулам, различаются незначительно.

5. Интегрирование методом Монте-Карло

Предположим, необходимо взять интеграл от некоторой функции. Воспользуемся неформальным геометрическим описанием интеграла и будем понимать его как площадь под графиком этой функции.

Для определения этой площади можно воспользоваться одним из обычных численных методов интегрирования: разбить отрезок на подотрезки, подсчитать площадь под графиком функции на каждом из них и сложить. Предположим, что для некоторой функции достаточно разбиения на 25 отрезков и, следовательно, вычисления 25 значений функции. Представим теперь, мы имеем дело с n-мерной функцией. Тогда нам необходимо 25^n отрезков и столько же вычислений значения функции. При размерности функции больше 10 задача становится огромной. Поскольку пространства большой размерности встречаются, в частности, в задачах теории струн, а также многих других физических задачах, где имеются системы со многими степенями свободы, необходимо иметь метод решения, вычислительная сложность которого бы не столь сильно зависела от размерности. Именно таким свойством обладает метод Монте-Карло.

Приложение. Справка: История метода.

Рождение метода Монте-Карло в Лос-Аламосе

Сначала Энрико Ферми в 1930-х годах в Италии, а затем Джон фон Нейман и Станислав Улам в 1940-х в Лос-Аламосе предположили, что можно использовать связь между стохастическими процессами и дифференциальными уравнениями «в обратную сторону». Они предложили использовать стохастический подход для аппроксимации мно-

гомерных интегралов в уравнениях переноса, возникших в связи с задачей о движении нейтрона в изотропной среде.

Идея была развита Уламом, который, по иронии судьбы, также, как и Фокс боролся с вынужденным бездельем во время выздоровления после болезни, и, раскладывая пасьянсы, задался вопросом, какова вероятность того, что пасьянс «сложится». Ему в голову пришла идея, что вместо того, чтобы использовать обычные для подобных задач соображения комбинаторики, можно просто поставить «эксперимент» большое число раз и, таким образом, подсчитав число удачных исходов, оценить их вероятность. Он же предложил использовать компьютеры для расчётов методом Монте-Карло.

Появление первых электронных компьютеров, которые могли с большой скоростью генерировать псевдослучайные числа, резко расширило круг задач, для решения которых стохастический подход оказался более эффективным, чем другие математические методы. После этого произошёл большой прорыв и метод Монте-Карло применялся во многих задачах, однако его использование не всегда было оправдано из-за большого количества вычислений, необходимых для получения ответа с заданной точностью.

Годом рождения метода Монте-Карло считается 1949 год, когда в свет выходит статья Метрополиса и Улама «Метод Монте-Карло». Название метода происходит от названия города в княжестве Монако, широко известного своими многочисленными казино, поскольку именно рулетка является одним из самых широко известных генераторов случайных чисел. Станислав Улам пишет в своей автобиографии «Приключения математика», что название было предложено Николасом Метрополисом в честь его дяди, который был азартным игроком.

Дальнейшее развитие и современность

В 1950-х годах метод использовался для расчётов при разработке водородной бомбы. Основные заслуги в развитии метода в это время принадлежат сотрудникам лабораторий ВВС США и корпорации RAND.

В 1970-х годах в новой области математики — теории вычислительной сложности было показано, что существует класс задач, сложность (количество вычислений, необходимых для получения точного ответа) которых растёт с размерностью задачи экспоненциально. Иногда можно, пожертвовав точностью, найти алгоритм, сложность которого растёт медленнее, но есть большое количество задач, для которого этого нельзя сделать (например, задача определения объёма выпуклого тела в n-мерном евклидовом пространстве) и метод Монте-Карло является единственной возможностью для получения достаточно точного ответа за приемлемое время.

В настоящее время основные усилия исследователей направлены на создание эффективных Монте-Карло алгоритмов различных физических, химических и социальных процессов для параллельных вычислительных систем.

Вопросы для самопроверки:

1. Что называется методом Монте-Карло?
2. Назовите типовые задачи метода Монте-Карло.
3. Какие вопросы исследует теория метода Монте-Карло?
4. Что называется разыгрыванием случайной величины?
5. Каким образом осуществляется разыгрывание дискретной случайной величины?
6. Каким образом осуществляется разыгрывание противоположных событий?
7. Каким образом осуществляется разыгрывание непрерывной случайной величины?
8. Опишите метод обратных функций.
9. Опишите метод суперпозиции.
10. Каким образом осуществляется приближенное разыгрывание нормальной случайной величины?

11. Каким образом осуществляется оценка погрешности метода Монте-Карло?

12. Расскажите историю метода Монте-Карло.

ЛЕКЦИЯ 10

АНАЛИЗ ДАННЫХ МЕТОДАМИ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Вопросы:

Введение.

1. Общая логическая схема статистического критерия;

2. Построение статистического критерия: принцип отношения правдоподобия;

3. Основные типы гипотез, проверяемых в ходе статистической обработки наблюдений.

Заключение.

Контрольные вопросы.

Введение

На разных стадиях формирования управленческих решений возникает необходимость в формулировке и экспериментальной проверке некоторых предположительных утверждений (гипотез) относительно величины или происхождения параметров анализируемой системы. Например, могут возникнуть предположения типа: "По данным наблюдений факторы экономической угрозы отсутствуют" или "Состояние контролируемого технического объекта изменилось".

Будем в дальнейшем обозначать высказанное нами предположение (гипотезу) буквой *H.*

Задача состоит в проверке непротиворечивости высказанной нами гипотезы имеющимся наблюдениям.

Процедура обоснованного сопоставления высказанной гипотезы с имеющимися выборочными данными (наблюдениями) $X_1, X_2,..., X_n$ осуществляется с помощью того или иного *статистического критерия* и называется статистической проверкой гипотез.

Результат подобного сопоставления может быть либо отрицательным (данные наблюдения противоречат высказанной гипотезе, и от этой гипотезы следует отказаться), либо неотрицательными (данные наблюдения не противоречат высказанной гипотезе, и, следовательно, ее можно принять в качестве одного из естественных и допустимых решений). При этом неотрицательный результат статистической проверки гипотез не означает, что высказанное предположительное утверждение является наилучшим, единственно подходящим: просто она не противоречит имеющимся наблюдениям. Однако такими же свойствами могут наряду с *H* обладать и другие гипотезы. Таким образом, даже статистически проверенное предположение *H* следует расценивать не как абсолютно достоверный факт, а лишь как достаточно правдоподобное, не противоречащее наблюдениям утверждение.

1. Общая логическая схема статистического критерия

На разных стадиях формирования решений возникает необходимость в формулировке и статистической проверке утверждений (гипотез) относительно природы или величины неизвестных параметров *объекта анализа* (ОА).

Например, необходимо установить факт наличия фактора системного сигнала в принимаемой смеси сигнал-шум или установить факт изменения состояния ОА по результатам последовательности наблюдений.

Процедура сопоставления истинности высказанной гипотезы с имеющимися выборочными данными $X_1, X_2, \ldots, X_n$ осуществляется с помощью *статистического критерия* и называется *статистической проверкой гипотез*.

В результате проверки гипотезы устанавливается либо отрицательный результат (данные противоречат высказанной гипотезе), либо неотрицательный (но не положительный!) - данные не противоречат гипотезе.

Неотрицательный результат не означает оптимальности решения.

Более того, он не означает, что утверждение верно.

Логическая схема статистической проверки гипотезы представляет собой следующую последовательность.

1. Выдвигается гипотеза H_o (например, ОА перешел из состояния исправности в состояние неисправности 1). В тех случаях, когда это возможно, формируется альтернатива (альтернативная гипотеза) H_1 (например, ОА не перешел в состояние неисправности 1).

2. Исходя из внешних (экзогенных) соображений, вытекающих из требований гиперсистемы (например, требований технического отдела), задаются:

- величиной уровня значимости критерия α (α - вероятность *ошибочного решения 1 рода* - отвергнута H_o, хотя на самом деле она верна, "пропуск цели", "пропуск неисправности");
- мощностью критерия $1-\beta$ (β - вероятность *ошибочного решения 2 рода* - принята H_o, хотя на самом деле она ошибочна, "ложная тревога"). β формируется лишь при наличии альтернативы H_1.

При ограниченном n α задается произвольно; при этом обычно задаются стандартные значения α=0,1; 0,05; 0,025; 0,01; 0,005; 0,001, в зависимости от значимости последствий, к которым приводит ошибочное неприятие Но;

при $n \to 0$ теоретически $\alpha, \beta \to 0$ для $\forall H_1$.

3. Задаются критической статистикой (функцией от X) $\theta_n = \theta(X_1, X_2, \ldots, X_n)$, подчиненной известному табулированному закону распределения $f(\theta_n)$ и определяющей меру расхождения $\mu(X_n, X \epsilon H_o)$.

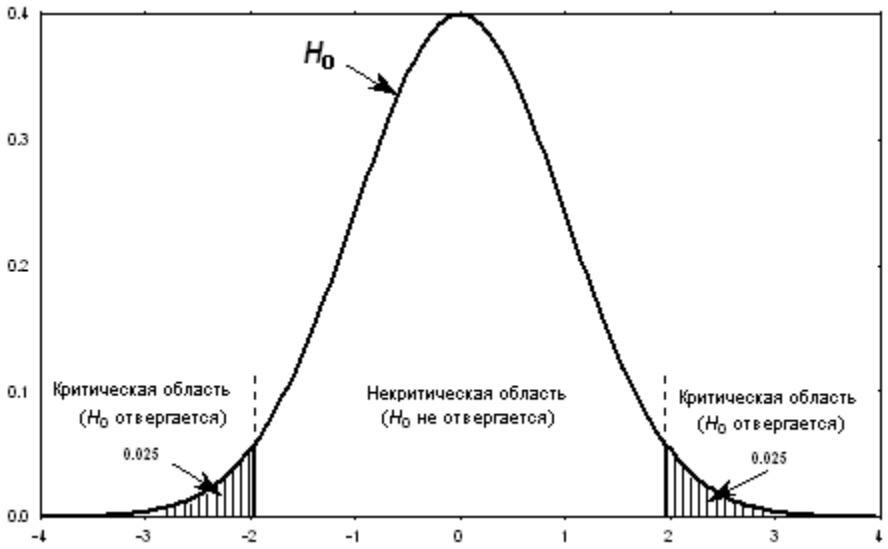


Рис. 1. Проверка гипотезы без альтернатив с двумя критическими областями

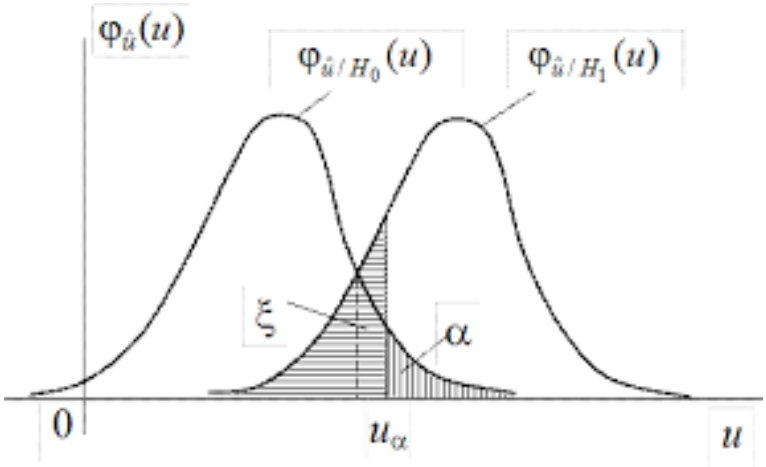


Рис. 2. Проверка гипотезы с альтернативой

Из таблиц $f(\theta_n, u)$ для заданного α определяют критические точки: $(1-\alpha/2)$-ая точка $\theta_{min}(\alpha/2)$ и $\alpha/2$-ая точка $\theta_{max}(\alpha/2)$, разделяющие область возможных значений θ на область принятия гипотезы H_o $\Gamma_o(\alpha)$ и область (двухсвязную) $\Gamma_1(\alpha)$, попадание в которую означает неприятие H_o.

4. Вычисляется $\theta_n = \theta(X_1, X_2, ..., X_n)$ и в зависимости от принадлежности $\theta_n \in \Gamma_o(\alpha)$ (H_o не противоречит наблюдениям $X_1, X_2, ..., X_n$) или $\theta_n \in \Gamma_1(\alpha)$ принимается или не принимается гипотеза H_o.

Гипотеза $H_o: x=x_o$ называется простой; во всех остальных случаях она называется сложной.

2. Построение статистического критерия. Принцип отношения правдоподобия

Рассмотрим общий принцип построения наилучших статистических критериев, т.е. критериев, обеспечивающих наибольшую мощность

$$1-\beta=max \qquad (1)$$

при заданном уровне значимости $\alpha \le \alpha_o$:

В качестве критической статистики используем отношение правдоподобия

$\theta_n = L_{H1}(X_1, X_2,..., X_n \mid \theta) / L_{H0}(X_1, X_2,..., X_n \mid \theta) =$

$= L(X_1, X_2,..., X_n \mid \theta_1) / L(X_1, X_2,..., X_n \mid \theta_o),$ (2)

где L_{H0}, L_{H1} - функции правдоподобия $X_1, X_2,..., X_n$, определенные для гипотез, соответственно,

$H_o: \theta = \theta_o$ и $H_1: \theta = \theta_1.$ (3)

Достаточно понятно, что чем больше наблюдения соответствуют H_o, тем больше будет значение $L(X_1, X_2,..., X_n \mid \theta_o)$ и тем меньше будет величина θ_n. Если $\theta_n(X_1, X_2,..., X_n) > \theta_\alpha$, где θ_α - α%-ая точка, определяемая из таблиц распределения $f(\theta_n)$ для уровня значимости α, то гипотеза H_o отвергается; в противном случае H_o не отвергается.

При этом вероятность ошибки "не признать" $\theta = \theta_o$ (в то время, когда он таковым является) равна $\int_{\theta_\alpha}^{\infty} a(\theta_m, z)dz = \alpha$

Известно, что в соответствии с леммой *Неймана-Пирсона*, *критерий отношения правдоподобия* (ОП) является наиболее мощным среди всех возможных критериев.

Пример. $L(X_1, X_2,..., X_n ; a_j, \sigma^2) =$

$= [(2\pi)^{n/2} \sigma^n]^{-1} \exp\{-(2\sigma^2)^{-1} \sum_{i=1}^{n} (X_i - a_j)^2\}, \quad j=0,1.$

Тогда критерий ОП имеет вид:

$\theta_n = L_{H1}/L_{Ho} = \exp\{-(2\sigma^2)^{-1} \sum_{i=1}^{n} (X_i - a_1)^2 - (X_i - a_o)^2\} \theta_\alpha,.$

Последнее выражение простым преобразованием приводится к соотношению

$\hat{\theta}_n \geq a_o + \Phi^{-1}(\theta_\alpha)\, \sigma / n^{1/2},$

где $\Phi^{-1}(\theta_\alpha)$ - θ_α-ая точка стандартного нормального распределения $N\{0, 1\}$, $\theta_{1-\alpha}$ - $(1-2\alpha)$ *% -ая точка.*

Получившееся правило проверки гипотезы не зависит от альтернативного значения параметра a_1 и поэтому верен для $\forall a_1 > a_o$. Такой критерий называется равномерно *наиболее мощным.*

В ряде случаев вместо критерия МП оказывается более удобным использовать критерий логарифма ОП:

$$\hat{\theta}_n = -2\ln\{\frac{L(X_1, X_2, ..., X_n \mid \theta_0)}{L(X_1, X_2, ..., X_n \mid \hat{\theta})}\}, \quad (4)$$

где $\hat{\theta}$ - МП-оценка параметра θ по выборке $X_1, X_2,..., X_n$.

При общих (для ММП) условиях регулярности $f\{X; \theta\}$ и $n \to \infty$ величина $\hat{\theta}_n \in \chi_n^2$ -распределению с n степенями свободы.

Заметим, что для нормального распределения оценка МП параметра θ_o представляет собой среднее значение $\overline{\theta}_n$. Тогда

$\theta_n = \sum_{i=1}^{n} \sigma^{-2}[\sum_{i=1}^{n} (X_i - \theta_o)^2 - \sum_{i=1}^{n} (X_i - \bar{\theta}_n)^2] = n \sigma^{-2}(X_i - \theta_o)^2$.

Поскольку $Xi \in N\{ \theta_o ; \sigma^2/n\}$, $\hat{\theta}_n \in \chi^2_n$.

3. Основные типы гипотез, проверяемых в ходе статистической обработки наблюдений

3.1. Гипотезы согласия. При обработке последовательностей наблюдений $X_1, X_2,..., X_n$ очень важно понять механизм их формирования, т.е. подобрать модельную функцию распределения $F_{mod}(X)$, адекватно описывающую истинное распределение F(X). Это приводит к задаче проверки гипотезы вида

$H_o: \quad F(X)=F_{mod}(X).$ (4)

$F_{mod}(X)$ может быть задана однозначно или с точностью до принадлежности некоторому параметрическому семейству $\{F_{mod}(X, x)\}$.

Проверка гипотез (4) осуществляется с помощью критериев *согласия* и опирается на некоторую меру $\mu[F_{mod}(X), \hat{F}_n0(X)]$ между гипотетическим и эмпирическим $\hat{F}_n(X)$ распределениями.

3.2. Гипотезы об однородности выборок наблюдений. Предположим, что ОА наблюдался в течении ***l*** сеансов (или l дней). Полученные наблюдения имеют вид рядов $(X_1, X_2,..., X_n)_1, (X_1, X_2,..., X_n)_2,...,(X_1, X_2,..., X_n)_l$.

Принятие решение о том, что ОА не изменил своего состояния, сводится к проверке одной из гипотез об однородности данных, имеющих вид:

$H_o: F_1(X_1)=F_2(X_2)=...=F_l(X_l);$

$H_o: a_1=a_2=...=a_l;$

$H_o: \quad \sigma_1^2= \sigma_2^2=...= \sigma_l^2$.

В случае отрицательного результата можно с заданным уровнем значимости утверждать, что состояние ОА не изменилось. Частный случай этой гипотезы при l=2 позволяет осуществить проверку аномальности одного или нескольких резко выделяющихся наблюдений.

3.3. Гипотезы о числовых значениях параметров исследуемой генеральной совокупности. Предположим, что в результате длительных наблюдений установлено среднее значение какого-то признака x, например, среднее число сообщений в сети a. Значимое отклонение от a означает возможность изменения состояния ОА. Для обнаружения этого изменения по наблюдениям $X_1, X_2,..., X_n$ осуществляется проверка статистической значимости гипотезы

$Ho: \quad E\{\hat{x}\}=a.$

Аналогично может проверяться значимость других предположений, например,

$H_o: \hat{r}\{\hat{x}_1, \hat{x}_2\}=0,$

где $\hat{r}\{\hat{x}_1, \hat{x}^2\}$ - выборочный коэффициент корреляции, построенный по двумерным наблюдениям $X_i =(X_1, X_2)_i, \quad i=1,...,n.$

К этому же классу задач относятся задачи проверки гипотез о параметрической стационарности и независимости рядов наблюдений.

3.4. Гипотезы о типе зависимости между компонентами исследуемого разведывательного признака.

С точки зрения задач управления большой интерес представляет характер зависимости между наблюдениями и параметрами состояния ОА или между различными признаками. Например, необходимо установить, как зависит среднее число самолетов в воздухе от интенсивности трафика авиационной радиосети связи. При этом проверяется гипотеза о виде этой зависимости, например,

H_o: $E\{x_2 | x_1\} = x_2 = b_0 + b_1 x_1$,

где b_0, b_1 - параметры модели.

Соответствующие статистические критерии называются критериями адекватности.

Заключение

1. Процедура обоснованного сопоставления предположительного утверждения (гипотезы) относительно природы или величины неизвестных параметров анализируемой системы с имеющимися в распоряжении результатами наблюдений осуществляется с помощью того или иного статистического критерия и называется статистической проверкой гипотез.

2. По своему прикладному содержанию высказываемые в ходе статистической обработки данных гипотезы подразделяют на следующие типы:

об общем виде закона распределения исследуемой случайной величины;

об однородности двух или нескольких обрабатываемых выборок;

о числовых значениях параметров исследуемой генеральной совокупности;

об общем виде зависимости, существующей между компонентами исследуемого многомерного признака;

о независимости и стационарности ряда наблюдений.

3. Все статистические критерии строятся по общей логической схеме. Построить статистический критерий - это значит:

а) определить тип проверяемой гипотезы;

б) предложить и обосновать конкретный вид функции от результатов наблюдения (критической статистики $\theta^{(n)}$, на основании значений которой принимается окончательное решение;

в) указать такой способ выделения из области возможных значений критической статистики $\theta^{(n)}$ области $\Gamma_n(H_1)$ отклонения проверяемой гипотезы H_o, чтобы было соблюдено требование к величине ошибочного отклонения гипотезы H 4o 0 (т.е. к уровню значимости критерия α).

4. "Качество" статистического критерия характеризуется уровнем значимости α, мощностью 1- β, свойствами несмещенности и состоятельности. В состоятельных критериях можно добиваться сколько угодно малых величин ошибок первого и второго рода (α и 7b 0) лишь за счет увеличения объема выборки n, на основании которой принимается решение.

При фиксированном объеме выборки можно делать сколь угодно малой лишь одну из ошибок (α или β), что сопряжено с неизбежным увеличением другой.

Вопросы для самопроверки:

1. Перечислите основные типы гипотез, проверяемых в ходе статистической обработки измерений;

2. В чем сущность гипотезы согласия? Гипотезы об однородности выборок наблюдений? Гипотезы о числовых значениях параметров исследуемой генеральной совокупности? Гипотезы о типе зависимости между компонентами исследуемого признака?

3. Что определяют уровень значимости и мощность статистических критериев?

4. Сформулируйте критерии проверки гипотез о параметрической стационарности и независимости рядов наблюдений.

5. Разработайте последовательность проверки гипотезы о равенстве сред-

них H_o: $E\{X_10\}=E\{X_2\}$ для альтернатив H_1: $E\{X_1\}>E\{X_2\}$ и H_1: $E\{X_1\}<E\{X_2\}$.

6. Как соотносятся уровень значимости критерия равенства средних α и значение табулированной t-статистики?

7. В чем состоит содержание гипотезы согласия?

8. В чем состоит содержание гипотезы об однородности выборок наблюдений?

9. В чем состоит содержание гипотезы о числовых значениях параметров исследуемой генеральной совокупности?

10. В чем состоит содержание гипотезы о типе зависимости между компонентами исследуемого признака?

ЛЕКЦИЯ 11

РЕГРЕССИОННЫЙ АНАЛИЗ И МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Вопросы:
1. Постановка задачи;
2. Простейшая модель линейной регрессии;
3. Линейная регрессия с несколькими переменными: Матричная форма

1. Общая постановка задачи восстановления зависимостей на основе метода наименьших квадратов. Предположим, что между двумя взаимосвязанными переменными которыми существует неизвестная исследователю непрерывная зависимость вида

$$Y = f(X, a) \in C^0,$$

Которую необходимо определить по результатам совокупности наблюдений

$$\{X_i, \quad Z_i = Y_i + v_i\}, \quad i = 1,..., n.$$

Здесь *a* – вектор параметров искомой зависимости, $\{v_i\}, \quad i = 1,..., n$ - вектор погрешностей измерений, C^0 - класс непрерывных функций.

В частности, если независимая переменная представляет собой время, то имеем задачу определения движения

$$Y = f(T, a).$$

Предположим, что с помощью какого-то метода удалось восстановить эту зависимость, т.е. получить ее оценку

$$\hat{Y} = \hat{f}(X, \hat{a}).$$

Найденную зависимость можно рассматривать, как модель исходной взаимосвязи. Естественно, желательно получить такую оценку, для которой априори выбранная метрика рассогласования между ней и исходной зависимостью была бы минимальной, т.е.

$$\mu(Y, \hat{Y}) = \mu\{f(X, a), \hat{f}(X, \hat{a})\} = \min .$$

Однако, поскольку истинные значения зависимости неизвестны, вместо них используются значения наблюдаемых измерений

$$\mu(Z, \hat{Y}) = \mu\{Z, \hat{f}(X, \hat{a})\} = \min$$

Если в качестве метрики рассогласования выбрать сумму квадратов разностей между наблюдениями и значениями модели, то получим *метод наименьших квадратов* (МНК):

$$\hat{f}(X, \hat{a}): \quad \sum_{i=1}^{n} (Z_i - \hat{Y}_i)^2 = \min$$

МНК был независимо разработан французским математиком Лежандром и немецким математиком К.Ф. Гауссом. Впервые Гаусс использовал МНК в 1799г. для определения движения астероида. Термин «регрессия» введен Френсисом Гальтоном для объяснения одного биологического процесса. Отсюда задача восстановления зависимостей по результатам наблюдений получила наименование регрессионного анализа.

Заметим, что выбор иной меры подобия приведет к другим вычислительным методам. Так, например, если в качестве меры подобия использовать сумму модулей

$$\hat{f}(X, \hat{a}): \quad \sum_{i=1}^{n} \setminus Z_i - \hat{Y}_i \mid = \min$$

то получим метод наименьших модулей.

Задача оценки зависимости (или задача построения математической модели зависимости) при выбранном критерии близости обычно решается итерационно в два этапа (рис. 1).

Рис. 1. Этапы решения задачи восстановления зависимости

На первом этапе, исходя из общих представлений выбирается структура модели. Например, если процесс носит сезонный характер, то в качестве структуры выбирают синусоидальную функцию или ряд Фурье

$$Y = \frac{c_0}{2} + \sum_{i=1}^{n} (a_i \sin(w_i X) + b_i \cos(w_i X).$$

При наличии апериодических процессов часто используют полиномиальные ряды

$$Y(X, a) = P_n(X, a) = a_0 + \sum_{i=1}^{n} a_i x^i$$

и т.п.

Заметим, что полиномиальные ряды обладают очень высоким уровнем общности. В частности, в соответствии с *аппроксимационной теоремой Вейерштрасса* для любой непрерывной функции $f(x) \in C^0$ на $x \in [a, b]$ отрезке можно подобрать последовательность многочленов P_n, равномерно сходящихся к этой функции на отрезке, т.е. $P_n(x) \xrightarrow[n \to \infty]{} f(x)$.

На втором этапе осуществляется оптимизационная оценка вектора параметров модели ***a*** в соответствии с выбранным критерием подобия. В частности, при использовании МНК, искомый вектор параметров определяется из условия

$$\hat{a}: \quad \sum_{i=1}^{n} (Z_i - \hat{f}(X_i, \hat{a}))^2 = \min$$

В случае, если и для оптимальных по выбранному критерию значений параметров найденная модель не удовлетворяет пользователя, осуществляется повторный выбор структуры модели и реализуется новая итерация.

В отношении погрешностей (или шумов) наблюдений $\{v_i\}, \quad i = 1,..., n$ обычно вводятся дополнительные ограничения:

1. Шумы наблюдений образуют независимую случайную последовательность $\operatorname{cov}\{v_i, v_j\} = 0, \quad \forall i, j = 1,..., n, \quad i \neq j;$

2. Наблюдения являются несмещенными, т.е. $E\{Z_i\} = Y_i, \quad \forall i = 1,..., n;$

3. Независимые переменные не являются случайными величинами, т.е. $\operatorname{cov}\{X_i, v_i\} = 0, \quad \forall i = 1,..., n;$

4. Для ряда наблюдений выполняется условие *гомоскедастичности*, т.е. $\operatorname{cov}\{v_i, v_i\} = \sigma^2\{Z_i\} = \sigma^2, \quad \forall i = 1,..., n.$

Во многих практических случаях в качестве дополнительного предположения используется гипотеза о гауссовском распределении погрешностей измерений, т.е. $v \in N\{0, \quad \sigma^2\}$.

В соответствие с теоремой Гаусса, выполнение перечисленных ограничений делает оценки по МНК наилучшими в классе всех линейных оценок.

2. Простейшая модель линейной регрессии. В рамках перечисленных выше ограничений рассмотрим простейший вариант задачи линейной регрессии с моделью наблюдений вида

$$Z_i = a_0 + a_1 X_i + v_i, \quad i = 1,..., n.$$

В соответствии с МНК ищем оценки параметров $\hat{a}_0$, $\hat{a}_1$, минимизирующих величину

$$S = \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} (Z_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Z_i - \hat{a}_0 - \hat{a}_1 X_i)^2$$

Находим экстремум.

$$\frac{\partial S}{\partial \hat{a}_0} = -2\sum_{i=1}^{n} (Z_i - \hat{a}_0 - \hat{a}_1 X_i) = 0;$$

$$\frac{\partial S}{\partial \hat{a}_1} = -2\sum_{i=1}^{n} X_i (Z_i - \hat{a}_0 - \hat{a}_1 X_i) = 0.$$

После приведения подобных членов получаем систему нормальных уравнений:

$$\sum Z_i = n\hat{a}_0 + \hat{a}_1 \sum X_i ;$$

$$\sum X_i Z_i = \hat{a}_0 \Sigma X_i + \hat{a}_1 \sum X_i^2 .$$

В матричной форме имеем

$$\begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix} \cdot \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \end{bmatrix} = \begin{bmatrix} \Sigma Z_i \\ \Sigma X_i Z_i \end{bmatrix}.$$

Соответствующее решение имеет вид:

$$\begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \end{bmatrix} = \begin{bmatrix} n & \Sigma X \\ \Sigma X & \Sigma X^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma Y \\ \Sigma XZ \end{bmatrix}. \qquad (1)$$

Заметим, что

$$\begin{bmatrix} n & \Sigma X \\ \Sigma X & \Sigma X^2 \end{bmatrix}^{-1} = \frac{1}{n\Sigma X^2 - (\Sigma X)^2} \begin{bmatrix} \Sigma X^2 & -\Sigma X \\ -\Sigma X & n \end{bmatrix},$$

отсюда $$\begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \end{bmatrix} = \frac{1}{n\Sigma X^2 - (\Sigma X)^2} \begin{bmatrix} \Sigma X^2 & -\Sigma X \\ -\Sigma X & n \end{bmatrix} \begin{bmatrix} \Sigma Z \\ \Sigma XZ \end{bmatrix}.$$

Следовательно,

$$\hat{a}_0 = \frac{\sum X^2 \sum Z - \sum X \sum XZ}{n\sum X^2 - (\sum X)^2};\ \hat{a}_1 = \frac{n\sum XZ - \sum X \sum XZ}{n\sum X^2 - (\sum X)^2}. \quad (2)$$

Введя соотношения центрирования:

$$\sum x^2 = \sum (X - \overline{X})^2$$

$$\sum xz = \sum (X - \overline{X})(Z - \overline{Z}),$$

где $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ - выборочные средние,

можно привести последние соотношения к виду:

$$\hat{a}_1 = \frac{\sum xz}{\sum x^2}, \quad (3)$$

$$\hat{a}_0 = \overline{Z} - \hat{a}_1 \overline{X} \ . \quad (4)$$

3. Линейная регрессия с несколькими переменными: Матричная форма

Модель регрессии допускает обобщение на случай ***m*** независимых переменных:

$\overline{Z}_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \ldots + a_m X_{1m} + v_i$, i=1,…,n.

В случае одного единственного наблюдения (подобные ситуации часто бывают в экономике), последнее выражения сводится к виду

$Z_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \ldots + a_m X_{mi} + v_i$, *i=1,…,n.*

Соответственно,

$E\{Y_i\} = \mu = a_0 + a_1 X_{1i} + a_2 X_{2i} + \ldots + a_m X_{mi}$, i=1,…,n.

] $\hat{a}' = [\hat{a}_0, \hat{a}_1, ..., \hat{a}_m]$; X=$\begin{bmatrix} 1 & X_{11} & \ldots & X_{m1} \\ 1 & X_{21} & \ldots & X_{m2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{1n} & \ldots & X_{mn} \end{bmatrix}$

В соответствии с МНК будем минимизировать сумму квадратов ошибок

$S = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Z - X\hat{a})'(Z - X\hat{a})$

Раскроем скобки полученной квадратической формы

$$\sum_{i=1}^{n} \varepsilon_i^2 = Z'Z - Z'X\hat{a} - \hat{a}'X'Z + \hat{a}X'X\hat{a} = \ Z'Z - 2\hat{a}'X'Z + \hat{a}X'X\hat{a}\,.$$

Для минимизации найденного выражения приравняем нулю первые производные

$$\frac{\partial}{\partial \hat{a}} \sum_{i=1}^{n} \varepsilon_i^2 = -2X'Y + 2X'X\hat{a} = 0\,. \quad (33)$$

Тогда $X'Y = X'X\hat{a} \quad \Rightarrow \ \hat{a} = (X'X)^{-1}X'Y.$

Вопросы для самопроверки:

1. В чем состоит МНК?
2. Какие метрики могут использоваться для построения альтернатив к МНК?
3. Опишите общую схему восстановления зависимости между двумя переменными.
4. Назовите основные этапы решения задачи восстановления зависимости.
5. Каким образом формируется непараметрическая структура зависимости?
6. Что называется линейной регрессией?
7. Как описывается линейная регрессия с несколькими переменными в скалярной форме?
8. Приведите матричную форму линейной регрессии с несколькими переменными.
9. Приведите матричной выражение для оценки коэффициентов линейной регрессии.
10. Кто являются авторами МНК?
11. Назовите свойства оценок по МНК.
12. Сформулируйте аппроксимационную теорему Вейерштрасса.
13. Сформулируйте условие гетероскедастичности.

ЛЕКЦИЯ 12

ОСНОВЫ ТЕОРИИ КЛАССИФИКАЦИИ И РАСПОЗНАВАНИЯ ОБРАЗОВ

1. Классификация. Формализованная постановка.

Исторически первыми в рамках работ по созданию искусственного интеллекта стали методы классификации, получившие название «распознавания образов» (Pattern Recognition).

NB! Отметим, что термин «pattern», кроме значения «образ», имеет еще значение «модель, стиль», «режим», «закономерность», «образ действия». В теории распознавания образов этот термин употребляют в самом широком смысле, имея в виду, что «образ» – это некоторое структурированное приближенное (обратите внимание – приближенное!) описание (эскиз) изучаемого объекта, явления или процесса.

Идея классификации заключается в группировании однотипных объектов на основании *гипотезы компактности.*

Суть гипотезы компактности состоит в том, что реализации одного и того же образа отображаются в пространстве характеристик геометрически близко расположенными точками, т.е. образуют «компактные» сгустки. В идеальном случае эти сгустки, образованные проекциями различных образов, не пересекаются. Более простыми словами эту гипотезу можно изложить так: объекты, схожие между собой, образуют в пространстве характеристик компактные сгущения.

В формальном виде гипотеза компактности представляется следующим образом: пусть имеется два объекта, обладающие k+1 характеристиками, причем последняя характеристика (z) является качественной и представляет собой имя класса. Тогда при равенстве (или, строго говоря, различии на величину не более малой величины ε для количественных переменных) k характеристик объектов x и b совпадает и k+1-я характеристика z:

$$\mathbf{b}_1 \approx \mathbf{x}_1 \cap \mathbf{b}_2 \approx \mathbf{x}_2 \cap \ldots \cap \mathbf{b}_k \approx \mathbf{x}_k \Rightarrow \mathbf{z}_b = \mathbf{z}_k$$

Замечание. Разделение характеристик объекта на описывающие (1, …, k) и целевые (z) является в задачах анализа данных весьма условным. Мы можем включить целевую характеристику z в число описывающих, а в качестве целевой (неизвестной) выбрать любую другую характеристику (x_j). Если при этом реализации некоторого образа A = $\{a_1, a_2, \ldots, a_n\}$ сохраняют компактность в новом пространстве характеристик $\{x_1, x_2, \ldots, x_{k-2}, z, x_j\}$, а множество ($A$, b) компактно в пространстве характеристик (признаков) $\{x_1, x_2, \ldots, x_{k-1}, z\}$, то значение новой целевой переменной x_j будет эквивалентным (в применяемой р) значению реализаций образа A. В математической постановке – это задача регрессии.

Целевыми могут быть не одна, а несколько характеристик. В частности, гипотеза компактности позволяет решать не только задачу анализа, но и обратную задачу – синтеза, когда по имени класса (образа) A восстанавливаются (прогнозируются) наиболее правдоподобные значения характеристик некоторого объекта b также отнесенного к классу A.

Несколько позже стало развиваться и другое направление группирования объектов, основанное только на сходстве их характеристик, при абсолютном отсутствии каких-либо данных об их «классовой принадлежности» – *кластеризация.* По этой причине, основываясь на том, что при классификации

имена классов и их границы задаются априорно, ее стали называть распознаванием с учителем (с обучением), а кластеризацию – распознаванием без обучения (без учителя), автоматической кластеризацией, таксономией.

NB! Термин «таксономия» чаще всего применяют для *иерархической кластеризации.*

Ввиду различия подходов к группированию объектов, далее методы классификации и кластеризации рассматриваются раздельно.

Сходство и различие распознавания с обучением и без. И классификация, и кластеризация предназначены для решения одного класса задач – определения некоторой характеристики объекта, выраженной в номинальной шкале – имени класса, либо кластера.

Основное же различие состоит в том, что имена и другие характеристики классов задаются заранее (в этом, собственно, и состоит процесс обучения системы распознавания); кластеры же формируются в ходе группирования объектов. В ряде случаев кластеры

Среди других различий подходов к распознаванию следует выделить:

а) Применяемый математический аппарат. Для решения задачи классификации используются статистические (вероятностные), структурные и логические методы математики. В кластеризации используется только одна характеристика: сходство объектов, выражаемое, обычно, расстоянием между ними в заданном признаковом пространстве.

б) Классификация, как правило, однозначна. Иными словами, при заданных условиях и неизменном математическом аппарате решение (минимальное расстояние до прообраза класса, максимальная вероятность принадлежности к классу и т.п.). Результаты некоторых алгоритмов кластеризации зависят, к примеру, от выбора начальных центров кластеров.

Постановка задачи классификации. Обычно используется следующая математическая постановка задачи распознавания с учителем: задана некоторая предметная область (M) разбитая на классы (прообразы) A_i; $i = 1,2, \ldots k$ так, что

$$\bigcup_{i=1}^{k}\{A_i) = M; \quad A_i \bigcap A_j = \varnothing; \quad j \neq i$$

Каждый класс описан набором характеристик, позволяющих их различить: $\{X\} = \{x_1, x_2, \ldots, x_m\}$ – *признаками.*

NB! Совокупность признаков принято называть *словарем признаков*. Далее будет показано, что словарь признаков при решении некоторых задач классификации образует *признаковое пространство*.

Задается также некоторый объект b, называемый реализацией, также представленный набором характеристик $\{b_1, b_2, \ldots, b_m\}$. Требуется найти качественную характеристику (ρ) распознаваемого объекта b, совпадающую с именем одного из классов, такую, что выполняется условие гипотезы компактности. Таким образом, задача классификации сводится к задаче выявления отношений эквивалентности на множестве объектов.

NB! Совокупность математических методов определения сходства класса и реализации, на основании которых определяется ρ, называется *решающим правилом*.

2. Алгоритм типовой системы классификации

Формируется рабочий словарь признаков, включающий только те характеристики, которые имеются и у реализации, и у описания класса.

NB! Характеристика свойства объекта становится только тогда, когда с ее помощью определяются схожесть объектов!

Проводится поочередное сравнение характеристик реализации с каждым из *K* классов.

На основании заданных решающих правил принимается решение о принадлежности реализации к одному из классов.

Рис.1. Алгоритм типовой системы классификации

1.

NB! Необходимо отметить, что строгая математическая постановка задачи достаточно далека от реалий действительности. *Разбиение* признакового пространства на практике не всегда применимо.

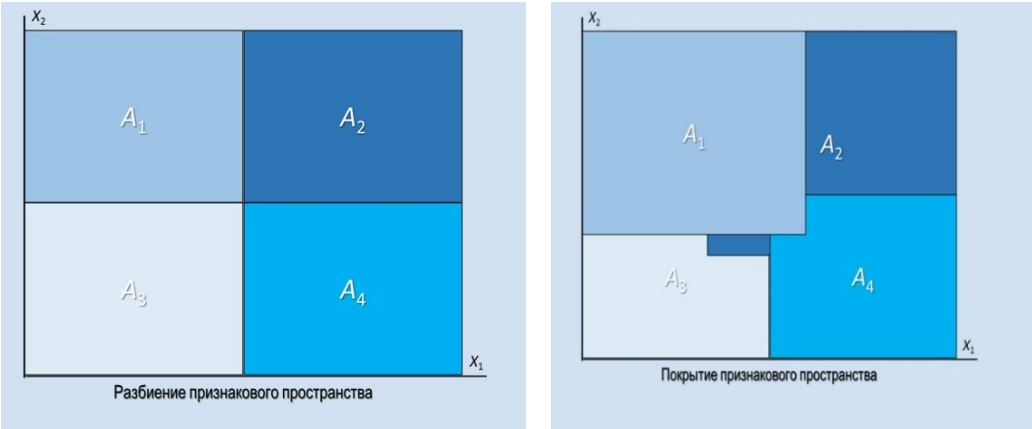


Рис.2-3. Разбиение и покрытие

Чаще области классов пересекаются в признаковом пространстве, что принято называть его *покрытием*. В этом случае решение о принадлежности объекта к одному из классов может носить вероятностный характер. Кроме того, возможны ситуации, когда *покрытие неполное*, то есть некоторые области признакового пространства не отнесены ни к одному из классов.

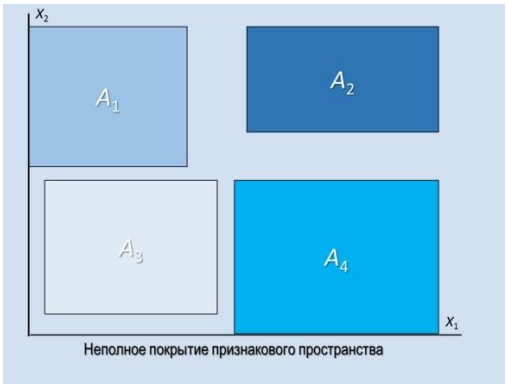


Рис. 4. Неполное покрытие

При попадании реализации именно в эту область компьютерные алгоритмы не в состоянии решить задачу – происходит отказ от распознавания. В этом случае требуется участие человека для уточнения границ классов (либо задания области нового класса) в признаковом пространстве. Очевидно, что в некоторых случаях возможно и *сочетание неполного покрытия с пересечением.* Обилием вариантов описания классов и их размещением в признаковом пространстве и обусловлено существование множества методов классификации.

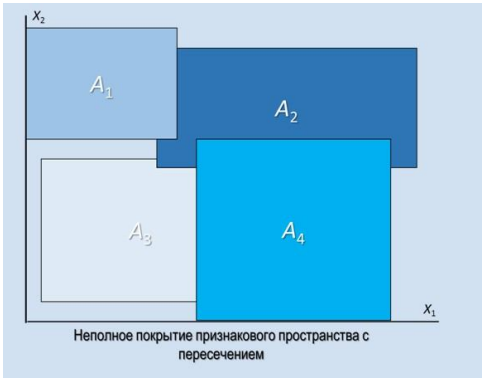


Рис. 5. Неполное покрытие с пересечением

3. Классификация методов распознавания с учителем

Несмотря на достаточно долгий период разработки методов распознавания, их общепринятой классификации не сложилось. Наиболее часто для этой цели используются следующие основания:

- *степень формализации решающих правил;*
- *способы описания классов объектов;*
- *используемый математический аппарат решающих правил;*
- *степень определенности результата.*

Далее при описании методов распознавания с учителем мы будем придерживаться данной классификации.

3.1. Классификация по степени формализации решающих правил

По данному основанию выделяют формальный и эвристический подходы.

Формальный подход предполагает построение математической модели предметной области, обеспечивающего применение математических методов для решения поставленной задачи. Рассматриваемые далее методы классификации относятся к формальным.

Эвристический подход основывается на попытке смоделировать опыт и интуицию человека. Примером такого подхода является использование для распознавания искусственных нейронных сетей.

NB! Методы формального подхода более общие, чем эвристического, однако, любая формализация предполагает некоторое «огрубление» исходных данных.

3.2. Классификация по способу описания классов объектов

По данному основанию выделяют интенсиональный и экстенсиональный подходы.

Суть *интенсионального* подхода заключается в замещении объектов класса некоторым абстрактным объектом – эталоном класса, либо, в некоторых случаях, определения признака как случайной величины с известным законом распределения, матрицей связи признаков и т.п.

Экстенсиональный подход, напротив, предполагает использование всех доступных объектов класса для определения принадлежности к одному из классов. Этот подход оперирует понятием «прототипа», которым может быть любой объект или их группа.

NB! При исследовании динамики объектов термин «прототип» принято заменять на «прецедент».

NB! Необходимо подчеркнуть, что в основе этого разделения лежат фундаментальные закономерности человеческого познания! Так, полагается, что для правого полушария человеческого мозга целостное представление окружающего мира, то для левого – выделение закономерностей, отражающих связи атрибутов объектов окружающего мира. Очевидный вывод из изложенного состоит в том, что для успешного решения задач (хотя мы об этом и не задумываемся) необходимо совместная работа обоих полушарий.

3.3. Классификация по используемому математическому аппарату

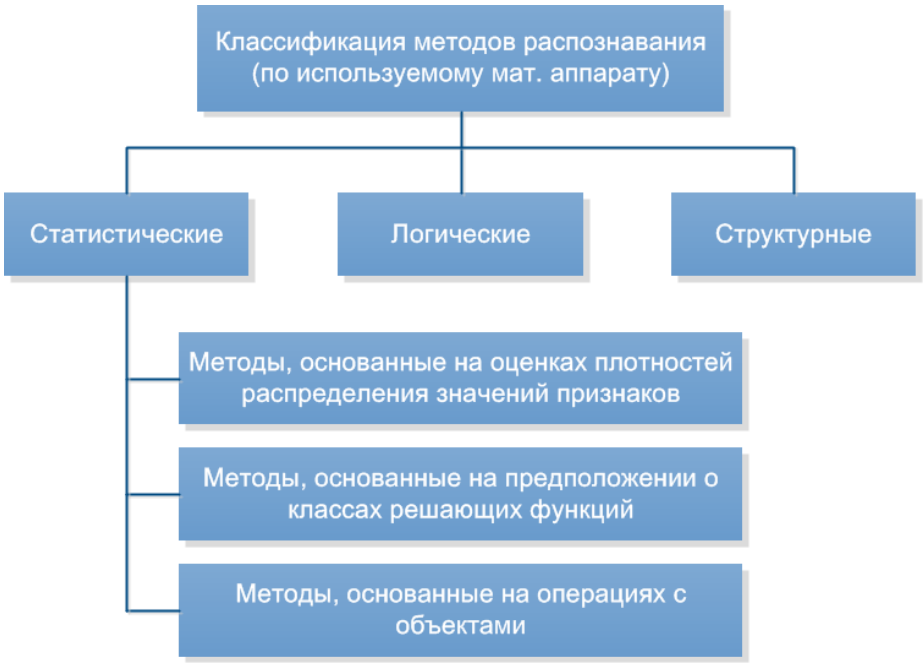


Рис.5. *Классификация по* используемому математическому аппарату

4. Методы классификации

Статистические методы. Группа статистических методов наиболее многочисленна, что обусловлено достаточно длительным периодом развития статистики как науки.

Первоначально данные методы развивались в русле классической параметрической статистики, для которой характерно использование интенсиональных методов.

К ним относятся методы, основанные на оценках плотностей распределения значений признаков. Нетрудно видеть, что данные методы предполагают знание закона распределения признака как случайной величины (в приведенном примере – нормального). Эти методы относятся к вероятностным. Для них разработан достаточно серьезный математический аппарат принятия решения о принадлежности реализации к одному из классов.

NB! Для данных методов достаточно существенны ограничения на применение центральной предельной теоремы теории вероятностей. Тем не менее, при наличии достаточно большого объема количественных данных необходимо провести проверку возможности их аппроксимации каким-либо известным законом распределения.

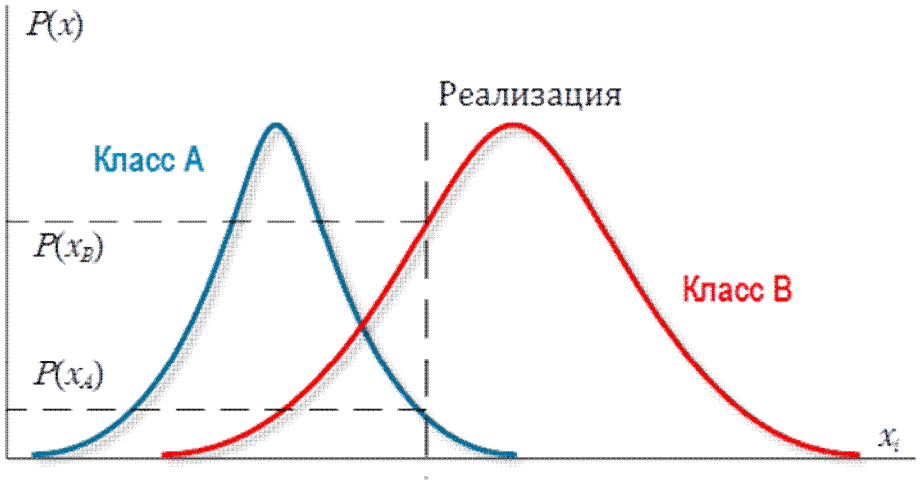


Рис.6. Классификация методом оценки вероятностей

Идея *методов, основанных на предположении о классах решающих функций* состоит в разделении признакового пространства на области, каждая из которых соответствует только одному классу. В простейшем случае – это линейная разделяющая поверхность (на случай многомерного пространства – гиперплоскость).

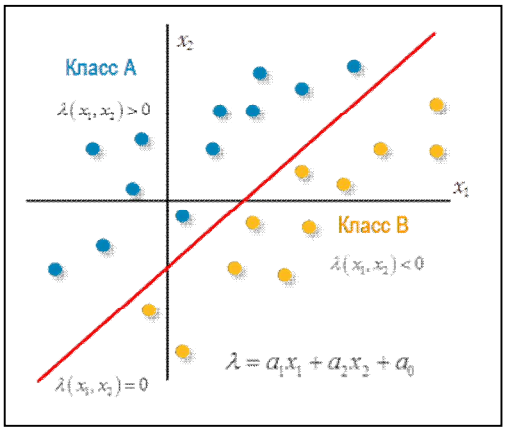


Рис.6. Классификация методом разделяющей

NB! Процедуры построения линейных решающих функций разработаны в 50-х годах XX века Ф. Розенблатом для использования в перпцептроне (устройства для распознавания изображений).

Один из способов, расширяющий область применения решающих функций, заключается в использовании кусочно-линейных функций, позволяющих разде-

лять достаточно сложные области. Другое расширение данного метода – ***метод опорных векторов*** (Support Vector Machine) – позволяет изменять геометрию признакового пространства с целью его разделения на области классов.

Методы, основанные на операциях с объектами, наиболее популярны в статистике объектов нечисловой природы. Кроме того, они применимы и при обработке количественных характеристик, особенно если неизвестен закон распределения признака как случайной величины. В основе рассматриваемых методов лежит ***метрик***а, рассматриваемая как величина, обратно пропорциональная сходству объектов.

Логические методы. Логические методы классификации основаны на установлении логических связей между признаками и классами объектов (в качестве последних также могут рассматриваться различные состояния одного и того же объекта).

Первоначально логические методы предусматривали построение высказываний булевой алгебры истинность или ложность которых при подстановке в них значений характеристик реализации свидетельствовала об ее принадлежности к одному из классов. При этом необязательно, чтобы характеристики были булевыми переменными.

Одним из современных направлений развития методов классификации, обеспечивающий подстановку значений характеристик реализации в высказывание вида «ЕСЛИ… И … ТО …» является математический аппарат *деревьев решений.*

Структурные методы. При структурном подходе к классификации объекты (прообразы) классов описываются не множеством характеристик, а структурой их элементов: каждый объект в таком представлении есть множество атомарных (или непроизводных) элементов, связанных определенными отношениями (соединенных между собой определенными способами). Это очень похоже на синтаксис (грамматику) естественных языков – любой объект может быть описан некоторым «предложением» формальной грамматики (существенно более простой относительно естественных языков).

Процедура классификации образа сводится к синтаксическому разбору «предложения», его описывающего, с целью определения корректности с точки зрения формальной грамматики, что эквивалентно проверке возможности порождения такого предложения в рамках фиксированного подмножества формальной грамматики, описывающей некоторый класс (прообраз).

NB! Методы формальной грамматики используются в языке описания изображений (Picture Definition Language, PDL).

Классификация по степени определенности результата. По данному основанию выделяются вероятностные и детерминированные методы. К вероятностным относятся методы, основанные на оценке плотности распределения значений признаков (рис. 5) и некоторые другие, основанные на методах классической параметрической статистики.

Все прочие методы относятся к детерминированным, дающим однозначный ответ о принадлежности реализации. Тем не менее, некоторые из методов обеспечивают количественную оценку принятого решения, выражающуюся, к примеру, либо разностью расстояний до объектов различных классов, либо «степенью уве-

ренности» – количеством присутствия данной градации признака у объектов различных классов.

Вопросы для самопроверки:

1. В чем состоит гипотеза компактности?
2. Чем отличаются обучение с учителем и без учителя?
3. Что называется кластеризацией?
4. Когда используется термин таксономия?
5. Что называется признаками?
6. Приведите формализованную постановку задачи классификации?
7. Что такое признаковое пространство?
8. Что называется решающим правилом?
9. Приведите алгоритм типовой системы классификации.
10. Что называется разбиением? Покрытием? неполным покрытием?
11. Приведите классификацию методов распознавания по степени формализации решающих правил.
12. Приведите классификацию методов распознавания по используемому математическому аппарату.
13. Как реализуются *методы классификации, основанные на оценках плотностей распределения значений признаков?*
14. Как реализуются логические *методы классификации?*
15. Перечислите методы оценивания информативности признаков

ЛЕКЦИЯ 13

ОЦЕНИВАНИЕ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ

Для оценки способности признаков разделять объекты, относящиеся к различным классам существует множество методов. Перечислим некоторые из них:

– метод сравнения расстояний;
– метод накопленных частот;
– энтропийный подход.

Метод сравнения расстояний

Метод сравнения расстояний основан на буквальном понимании гипотезы компактности признаковое пространство полагается тем более лучшим, чем больше среднее межклассовое расстояние и меньше среднее внутриклассовое. В формальном виде это описывается следующими выражениями:

$$\mathbf{J} = \frac{\Delta(\mathbf{A_i}, \mathbf{A_j})}{\Omega_\mathbf{i} + \Omega_\mathbf{j}},$$

где Ω_k – средняя длина ребер графа, соединяющего все объекты k-го класса, определяется как

$$\Omega_k = \frac{1}{C_m^2} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d(a_i, a_j),$$

где m – количество объектов в классе; C– количество сочетаний, $d(a_i, a_j)$ – расстояния между парами объектов класса.

$d(A_i, A_j)$ – среднее расстояние между i-м и j-м классами:

$$d(A_i, A_j) = \frac{1}{m_i m_j} \sum_{k=1}^{m_i} \sum_{l=i}^{m_j} d(a_{ik}, a_{jl}).$$

Метод накопленных частот

Сущность метода накопленных частот состоит в следующем: для значений одного из признаков (x) каждого из классов строится гистограмма. Далее для каждого интервала подсчитываются накопленные частоты – количество объектов каждого класса от начала гистограммы до текущего значения включительно. Максимальная разность накопленных частот и будет оценкой информативности.

Следует отметить относительность полученной оценки. Насколько хороша измеренная ей информативность, можно судить только по сравнению с аналогичными расчетами для других признаков этих же классов.

Энтропийный подход

В терминах теории информации мерой информативности признака служит энтропия (H) распределений плотности вероятности прообразов (классов). Суть методов заключается в определении изменения энтропии как меры неопределенности до и после измерения признака. Наиболее распространенными методами оценки информативности в этом подходе являются:

– метод Шеннона;
– метод Джини;
– метод Кульбака.

Метод Шеннона.

Рассмотрим метод Шеннона на следующем примере.

Пусть распределения объектов *k* классов проецируются на ось признака *x* с точностью до *t* градаций. Вероятность попадания объектов *i*-го образа в *j*-ю градацию обозначим *P*(*j*/*i*) (очевидно, что она определяется как отношение объектов *i*-го класса, попавшим в *j*-ю градацию к общему количеству объектов в нем).

Просуммировав для j-й градации вероятности всех попадания в нее объектов всех *k* классов, получим величину

$$P_j = \sum_{i=1}^{k} P_{j|i}$$

Вклад i-го класса в эту сумму определяется как

$$r_j = \frac{P_{j|i}}{P_j}$$

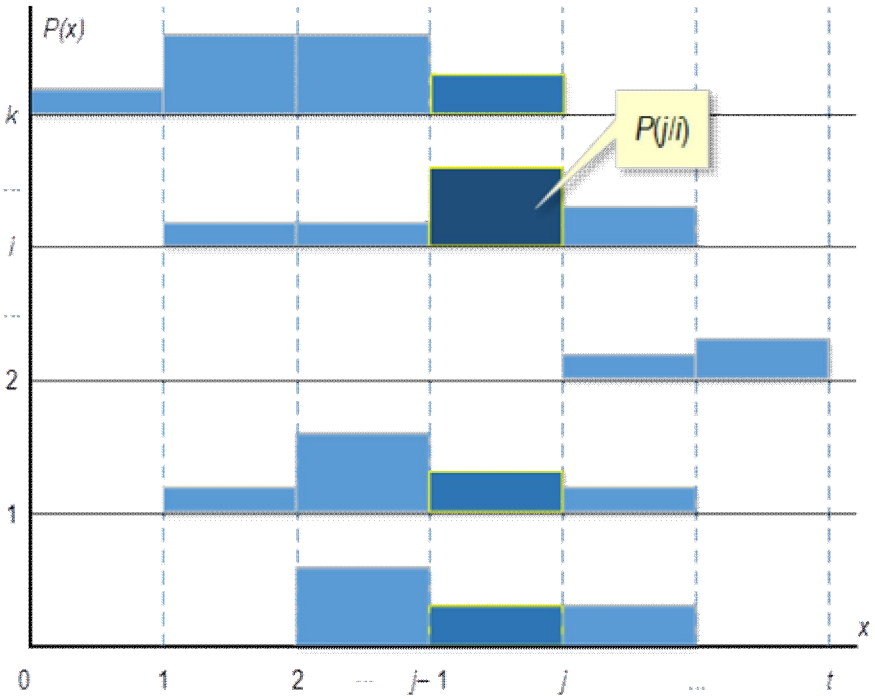


Рис.7. Пример расчета информативности по Шеннону

В соответствии с положениями теории информации энтропия *j*-й градации выражается как

$$H_j = \sum_{i=1}^{k} r_i \log r_i .$$

Из свойства аддитивности энтропии следует, что общая неопределенность классификации по признаку *x* имеет вид

$$H_x = \sum_{i=1}^{k} H_i P_i .$$

Если исходная неопределенность H_0 ситуации равнялась log *k*, то количество информации I_x, получаемой в результате измерения признака *x*, равна $H_0 - H_x$.

Метод Джини

Показатель (индекс) Джини позволяет определить информативность разбиения объектов обучающей выборки по некоторому правилу. При этом узлов

разбиения может быть только два, а классов в обучающей выборке – произвольное количество. Смысл данного индекса в буквально соответствует *гипотезе компактности*: разбиение тем лучше, чем оно «чище», то есть, если в нем содержатся объекты, принадлежащие только одному классу. Метод Джини, по существу, оценивает количество «мусора» в разбиении, поэтому, чем меньше его значение, тем выше качество разбиения.

Количественная оценка качества разбиения множества T на два подмножества T_1 и T_2, содержащих соответственно по N_1 и N_2 элементов, по методу Джини определяется следующим образом:

$$G(S,T) = \frac{N_1}{N} G(T_1) + \frac{N_2}{N} G(T_2),$$

где $G(T_i)$ – собственно индекс Джини – определяется как

$$G(T_i) = 1 - \sum_{j=1}^{K} P_j^2 .$$

Здесь K – количество классов в выборке; P_j – частота вхождения j-го класса в подмножество.

Метод Кульбака

Показатель информативности признака, получивший название *дивергенция Кульбака*, вычисляется по следующей формуле:

$$I(x) = \sum_{j=1}^{t} [P_{j1} - P_{j2}] \log_2 \frac{P_{j1}}{P_{j2}},$$

где t – количество градаций признака.

Вероятность появления j-й градации в i-м классе определяется как

$$P_{ji} = \frac{m_{ji}}{card\ A_i}$$

Здесь m_{ji} – частота появления j-й градации в i-м классе, а $card(A_i)$ – мощность множества элементов, отнесенных к i-му классу, то есть количество объектов данного класса в обучающей выборке.

NB! Заметьте: дивергенция Кульбака несимметрична, хотя, ее иногда называют расстоянием. Ее значение будет зависеть от того, какой из классов назначен первым. На практике первым классом назначается некоторая эталонная модель, сравниваемая с экспериментальными данными, представляющими второй класс объектов.

Сравнительный анализ методов

Метод сравнения расстояний в рассмотренном множестве – единственный, позволяющий оценить за один «проход» информативность рабочего словаря признаков, при этом он не требует разбиения области значений признака на градации. Все остальные методы оценивают одновременно информативность только одного признака и предполагают введение градаций. При этом только метод Шеннона и Джини дают абсолютную оценку информативности. Прочие методы позволяют говорить об оценке только в относительном плане – более высокая или более низкая. Так, для метода накопленных частот оценка информативности одного признака сопоставима только с оценкой другого признака, либо для других примеров обучающей выборки аналогичного объема и состава. Обязательным условием для данного метода, кроме того, является равное количество объектов различных классов в обучающей выборке.

Наконец только методы Шеннона и Джини позволяют оценивать информативность для произвольного количества классов.

NB! Следует отметить, что метод накопленных частот является наиболее простым для вычислений.

Сравнительный анализ методов приведен в таблице.

Метод	Предмет оценивания	Зависимость от способа кодирования	Кол-во сравниваемых классов	Зависимость от объема выборки	Оценка информативности	Необходимость введения градаций признака
Сравнения расстояний	пространство	да	2	нет	относит.	нет
Накопленных частот	признак	да	2	равное для классов	относит.	есть
Энтропийные методы						
Шеннона	признак	нет	любое	нет	абсол.	есть
Джини	признак	нет	любое	нет	абсол.	есть
Кульбака	признак	нет	2	нет	относит.	есть

NB! Следует, наконец, пояснить необходимость оценки информативности признаков. Первоначально актуальность этой задачи обосновывалась необходимостью экономии ресурсов вычислительной техники: чем меньше признаков обрабатывалось совместно, тем быстрее решалась задача классификации.

В настоящее время необходимость формирования эффективного признакового пространства обусловлена тем, что низкоинформативные признаки, как правило, вносят в процесс распознавания избыточный уровень «шумов», что негативно сказывается на результате. Поэтому одной из задач реализации процесса распознавания с учителем является подбор эффективного признакового пространства.

Необходимо отметить, что информативность отдельного признака может быть гораздо более низкой, чем их совокупности. В этом случае принято говорить о зависимых признаках. Рассмотренные ранее способы оценивания информативности признаков или признакового пространства в целом предполагают независимость признаков.

Строгих математических методов, позволяющих выделить среди множества зависимых признаков подмножество достаточно информативных, кроме прямого перебора, не существует. Однако такая операция весьма трудоемка даже для современных компьютеров.

Вопросы для самопроверки:

1. Перечислите методы способности признаков разделять объекты.
2. В чем состоит метод сравнения расстояний?
3. В чем состоит метод накопленных частот?
4. Назовите наиболее распространенные методы оценки информативности в энтропийном подходе.
5. Опишите метод Шеннона.
6. Что определяет индекс Джини?
7. Что называется дивергенцией Кульбака?
8. Сравните методы анализа информативности признаков.

ЛЕКЦИЯ 14

МЕТОДЫ КЛАССИФИКАЦИИ

Основываясь на возможности распространения концепции расстояния на все области статистики, следует пояснить различия интенсионального и экстенсионального подходов к описанию объектов прообраза (класса).

В интенсиональном подходе при использовании концепции расстояния множество образов, относящихся к одному прообразу (классу), заменяется некоторым абстрактным объектом, обладающим усредненными характеристиками всех объектов данного класса – *эталоном* класса.

В экстенсиональном подходе важную роль играет *прототип* – ближайший к реализации объект класса.

В наиболее простых методах распознавания с учителем решение об отнесении реализации к одному из образов принимается по критерию минимального расстояния между реализацией и эталоном класса или прототипом, как это показано на рис. 10.

2.

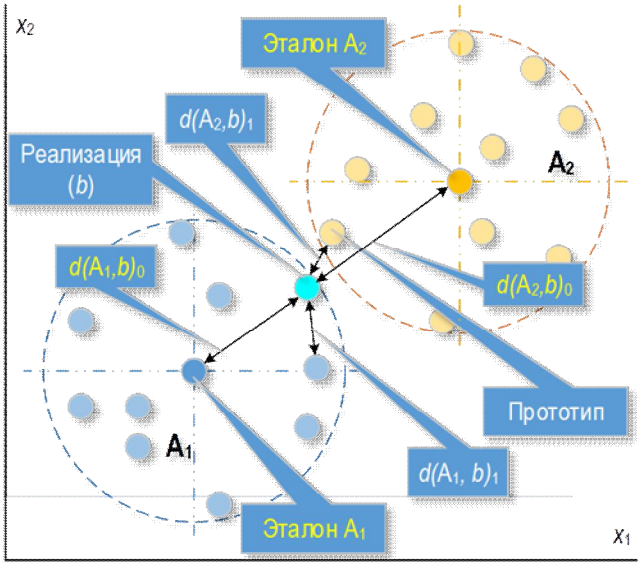


Рис.10. Критерий минимального расстояния

NB! Недостатком рассмотренных методов является высокая зависимость результата от распределения прообразов в признаковом пространстве, а также высокая чувствительность к «выбросам» – аномальным объектам класса. Несколько снизить влияние аномалий позволяет описанный ниже метод.

1. Метод k-ближайших соседей

Суть метода состоит в том, что решение об отнесении реализации (*b*) к одному из классов принимается не по одному примеру (эталону или прототипу), а по некоторому множеству наиболее близких к ней объектов, что позволяет несколько сгладить влияние выбросов. Для этого вокруг реализации строится область (окружность, а в многомерном пространстве – гиперсфера), в которую попадают «голосующие» объекты обучающей выборки, принадлежащие различным классам. Радиус гиперсферы выбирается таким, чтобы в нее попали *k* объектов обучающей

выборки (отсюда и название метода). Его значение, точнее, значение k, зависит от объема обучающей выборки (N): обычно полагают $k \approx N^{1/2}$, $k \approx \log_2(N)$ и т.п.

Основной проблемой метода k-ближайших соседей является принятие решения при четном k и равном представительстве объектов различных классов в выделенной области. В этом случае значение k либо увеличивается, либо уменьшается на единицу.

NB! Здесь и далее принято, что размерность алфавита классов равна двум. На практике же их может быть и больше. Однако здесь можно воспользоваться принципом «здорового шовинизма», полагая, что алфавит классов содержит «нужный класс» и класс «все остальные».

2. Метод дробящихся эталонов

Другим методом, в котором приняты меры для нейтрализации неоднородного распределения объектов различных классов в признаковом пространстве, является метод «дробящихся эталонов», разработанный Н.Г. Загоруйко. Идея метода состоит в последовательном уточнении эталонных описаний, в результате чего признаковое пространство разбивается на области, принадлежащие различным классам.

NB! На заключительных этапах дробления в признаковом пространстве могут оказаться области, не отнесенные ни к одному из классов. В этом случае для распознавания реализации может быть применен любой другой из рассмотренных выше методов.

3. 1R-правила

Достаточно тривиальным алгоритмом для формирования правил классификации объектов является так называемый 1R-алгоритм.

NB! Его название расшифровывается как «one rule» или в русскоязычном переводе – «1-правило».

Идея 1R-алгоритма (на примере номинальных переменных) состоит в следующем: для каждой градации переменной по известной обучающей выборке определяется достоверность проявления в каждом из классов. Пара «градация-класс» с наибольшей достоверностью и представляет собой 1-правило.

NB! Существенным недостатком 1R-алгоритма является сверхчувствительность (overfitting). Дело в том, что для уникальных значений достоверность определяется максимальной. Чтобы избежать этого, целесообразно ввести градации количественной переменной, как это и показано в примере.

Несмотря на простоту реализации, 1R-алгоритм во многих случаях обеспечивает достаточно надежное распознавание. Кроме того, немногочисленность полученных правил позволяет легко интерпретировать и использовать полученные результаты.

4. Метод потенциальных функций

Свое название данный метод получил в связи с использованием аналогии электрических потенциалов: объекты, относящиеся к одному классу, наделяются единичным зарядом (потенциалом) одинаковой полярности. Объекты другого

класса обладают точно таким же по значению потенциалом, но противоположной полярности.

NB! При размерности алфавита классов более двух можно выделить «нужный класс» и «прочие классы».

Из курса физики известно, что суммарный наведенный потенциал в некоторой точке определяется как сумма потенциалов всех зарядов, находящихся достаточно близко к этой точке:

$$K(x) = \sum_{i=1}^{N_1+N_2} K(x, x_i)$$

где N_1, N_2 - размеры обучающих выборок из первого и второго класса.

$K(x, x_j)$ называют *потенциальной функцией i-го заряда*. Она, как и электрический заряд, убывает с ростом расстояния между x и x_i.

Чаще всего в качестве потенциальной используется функция, имеющая максимум при $x = x_i$ и монотонно убывающая до нуля при $|x - x_i| \to \infty$.

Этим требованиям удовлетворяет, например, функция

$$K(x, x_i) = \exp(-\alpha^2 d^2(x, x_i)),$$

или функция

$$K(x, x_i) = (1 + \alpha^2 d^2(x, x_i))^{-1}$$

где a – параметр функции, определяющий скорость ее убывания; $d(x, x_i)$ – расстояние между точками x и x_i.

Изолинии, соединяющие точки с нулевым потенциалом, будут являться границами класса в признаковом пространстве. Впрочем, их построение необязательно: можно просто рассчитать кумулятивный потенциал только в той точке признакового пространства, которая соответствует реализации.

Графическая иллюстрация метода на примере одного признака приведена на рис. 11. Классификация неизвестного наблюдения осуществляется по знаку потенциальной функции в точке его расположения.

При этом ограничений на построение потенциальных функций в многомерном пространстве нет.

NB! Выбор параметра a достаточно неоднозначен. Так, если потенциальная функция убывает очень быстро, то можно добиться практически безошибочного разделения объектов обучающей выборки. Однако при этом могут возникнуть зоны неопределенности (там, где потенциал близок к нулю). При слишком «пологих» склонах функции может необоснованно увеличиться количество ошибок распознавания, в том числе и на обучающих объектах. Некоторые рекомендации в этом отношении можно получить, рассматривая метод потенциальных функций со статистических позиций (восстановление плотности распределения вероятностей $p(x)$ или разделяющей границы по выборке с использованием процедуры типа стохастической аппроксимации), однако, этот вопрос выходит за рамки учебника.

NB! В приведенном примере использован так называемый «наивный подход», основанный на суммировании потенциалов. Из примера видно, что потенциал класса А «нейтрализуется» потенциалом класса В. Особенно это заметно при количественном превосходстве объектов одного класса над другим. Для устране-

ния этого недостатка разработаны специальные методы, основанные на подборе весовых коэффициентов объектов обучающей выборки.

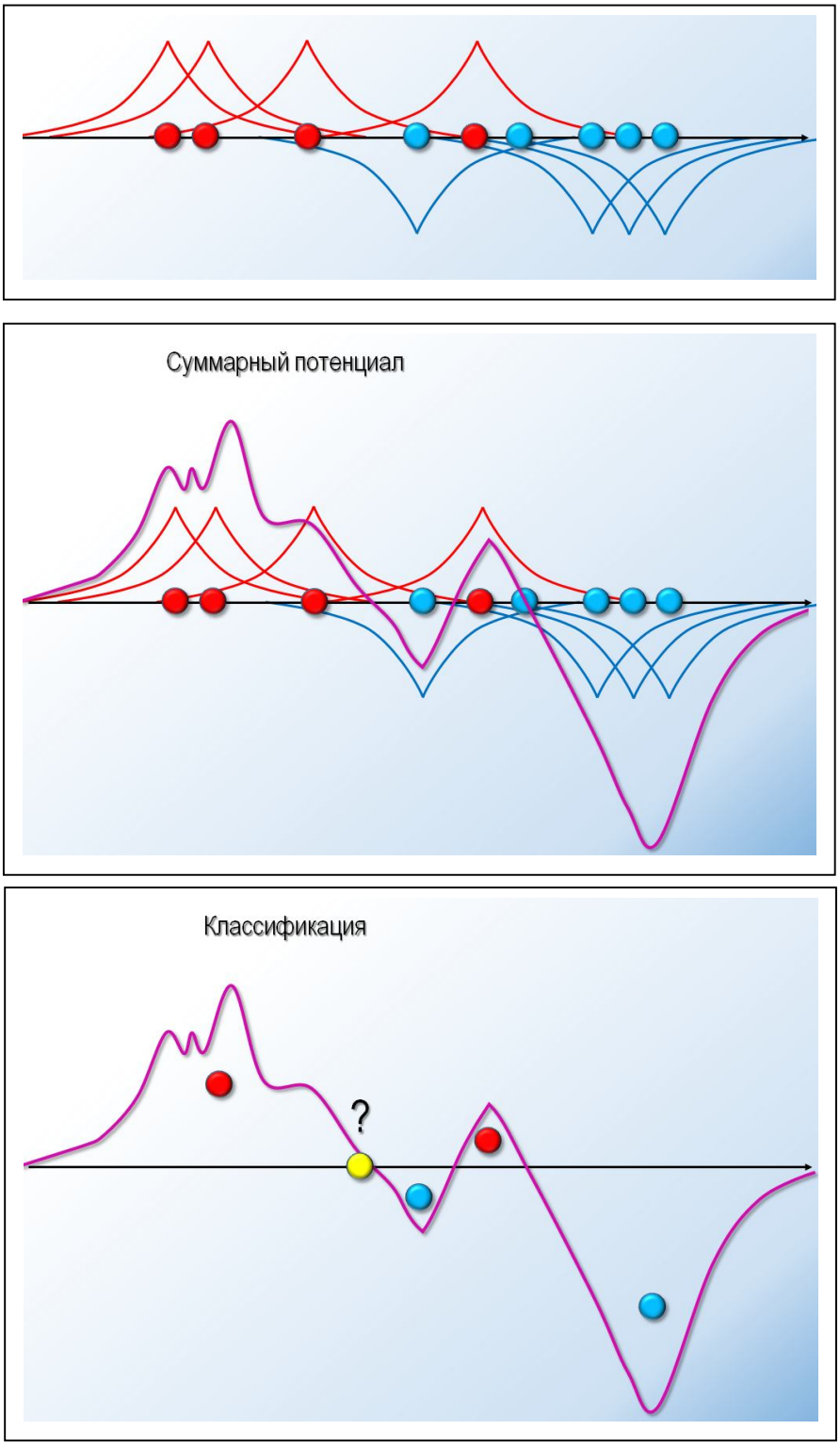


Рис.11. Метод потенциальных функций

5. Классификатор Байеса

В основе байесовского классификатора лежит теорема Байеса, с которой Вы уже знакомы из курса теории вероятностей:

$$P(H|B) = \frac{P(H)P(B|H)}{P(B)}$$

где $P(H)$ – априорная вероятность гипотезы H; $P(H|B)$ – апостериорная вероятность гипотезы H при наступлении события B; $P(B|H)$ – вероятность наступления события B при истинности гипотезы A; $P(B)$ – полная вероятность события B.

NB! «… теорема Байеса является основой теории вероятности, точно так же как и теорема Пифагора есть основа геометрии».

[Jeffreys, Harold (1973), Scientific Inference (3rd ed.), CambridgeUniversity Press, p. 31]

С приведенным определением трудно не согласиться: она позволяет переставить местами причину и следствие. Так, зная, с какой вероятностью причина приводит к некоторому событию, с помощь этой теоремы можно рассчитать вероятность того, что именно эта причина привела к проявлению наблюдаемого события.

NB! В научной литературе, особенно в англоязычной классификатор Байеса часто называют наивным (Naive Bayes). Это объясняется тем, что основным допущением при выводе теоремы Байеса является независимость событий. Однако, несмотря на наивность эти классификаторы дают достаточно хорошие результаты во многих жизненных ситуациях.

Вопросы для самопроверки:

1. Поясните различия интенсионального и экстенсионального подходов к описанию объектов прообраза.
2. В чем состоит метод k ближайших соседей?
3. Объясните содержание метода дробящихся эталонов.
4. В чем состоит алгоритм 1R-правила?
5. Что называется сверхчувствительностью (overfitting) алгоритма?
6. В чем состоит метод потенциальных функций?
7. Что называют *потенциальной функцией i-го заряда?*
8. Что является основой байесовского классификатора?

Литература:

1. Загоруйко Н.Г. Прикладной анализ данных и знаний. – Новосибирск : Изд-во НГУ, 1990.
2. Плэтт В. Информационная работа стратегической разведки: основные принципы. – М.: Изд-во иностр. лит-ры, 1958.
3. Паклин Н.Б., Орешков В.И., Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2013.
4. Барсегян А.А. и др. Анализ данных и процессов. – СПб: БХВ-Петербург, 2009.
5. Горелик А.Л., Скрипкин В.А. Методы распознавания :Учеб. пособие для вузов. – М.: Высш. шк., 2004.
6. Фу К. Структурные методы в распознавании образов. – М.: Мир, 1978.
7. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. – М.: Наука, 1970.

8. Лапко В.А. Непараметрические коллективы решающих правил. – М.: Наука, 2002.
9. Журавлев Ю.И., Камилов М.М., Туляганов Ш.Е. Алгоритмы вычисления оценок и их применение. – М.: Фан, 1989.
10. Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977.
11. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988.
12. Орлов А.И. Прикладная статистика. – М.: Экзамен, 2004.

ЛЕКЦИЯ 15
ДИСКРИМИНАНТНЫЙ АНАЛИЗ

1. Линейный дискриминантный анализ

Пусть теперь в *k*-мерном признаковом пространстве имеются два облака точек $\{X_i\}$ и $\{Y_j\}$, $i = 1,\ldots, n$, $j = 1,\ldots, m$. Они характеризуются своими центрами a_1 и a_2 и своими ковариационными матрицами Σ_1, Σ_2, вместо которых рассматривают соответствующие выборочные матрицы рассеяния $S_1 = (n-k)\hat{\Sigma}_1$ и $S_2 = (m-k)\hat{\Sigma}_2$. Обычно бывает выгодно с самого начала перейти к безразмерным величинам, поделив каждую компоненту на соответствующее ей СКО. Поставим задачу о выборе направления проектирования следующим образом: будем требовать, чтобы рассеяние внутри облаков - проекций было минимальным и при этом эти облака разошлись как можно дальше друг от друга.

Рассмотрим формальную постановку этой задачи. Пусть имеются две квадратичные формы C^TAC и C^TBC с симметричными положительно-определенными матрицами A и B. Требуется найти единичный вектор C, являющийся решением следующей оптимизационной задачи:

$$C^TAC \to \min; \quad C^TBC \to \max; \quad C^TC = 1.$$

Предположим, что для второй формы уже найдено удовлетворительное значение: $C^TBC=b$, $b>0$. Будем решать задачу

$$C^TAC \to \min; \quad C^TBC = b,$$

а условие нормировки $C^TC = 1$ учтем позднее. Составим функцию Лагранжа

$$L(C,\lambda) = C^TAC - \lambda(C^TBC - b)$$

и приравняем нулю ее производные по C и по λ:

$$\begin{cases} \dfrac{\partial L}{\partial C} = 2AC - \lambda \cdot 2BC = 0 \\ \dfrac{\partial L}{\partial \lambda} = C^TBC - b = 0 \end{cases} \Rightarrow \begin{cases} (A - \lambda B)C = 0 \\ C^TBC = b \end{cases} \Rightarrow \begin{cases} C^TAC = \lambda C^TBC \\ C^TBC = b \end{cases}.$$

Отсюда сразу следует, что при любом b искомый вектор C^* должен быть собственным вектором пучка квадратичных форм $\{A, B\}$, соответствующим одному из его собственных чисел λ*, при этом вспомним также об условии $C^{*T}C^* = 1$. Вычислим соответствующее значение функции Лагранжа или, что то же самое, значение минимизируемой функции:

$$L(C^*,\lambda^*) = C^{*T}AC^* = \lambda C^{*T}BC^* = \lambda b.$$

Таким образом, решение задачи состоит в том, что C – единичный собственный вектор пучка $\{A, B\}$, соответствующий его ***минимальному*** собственному числу λ. Это диктует два подхода к решению исходной задачи об оптимальном проектировании.

1. В качестве матрицы A выбираем сумму ковариационных матриц рассматриваемых совокупностей: $A = \Sigma_1+\Sigma_2$. Это означает, что минимизируется сумма внутренних дисперсий проекций $C^T\Sigma_1C + C^T\Sigma_2C$.

В качестве матрицы B выбираем произведение расстояний между центрами исходных облаков: $B = (a_1-a_2)(a_1-a_2)^T$, тогда $C^TBC=[(C^Ta_1C - C^Ta_2C)(C^Ta_1C - C^Ta_2C)^T] = (C^Ta_1C - C^Ta_2C)^2$ – квадрат расстояния между центрами облаков-проекций. Недостаток этого подхода в том, что матрица B имеет ранг 1, поэтому у пучка $\{A,B\}$ имеется единственное собственное подходящее собственное число λ. Данный подход обеспечивает проектирование только на прямую, что дает единственную ***дискриминантную компоненту***.

2. В качестве матрицы *A* снова выбираем сумму ковариационных матриц $\Sigma_1+\Sigma_2$, а в качестве *B* – ковариационную матрицу объединенного облака. При этом минимизируется сумма внутривыборочных дисперсий проекций и одновременно максимизируется их межвыборочная дисперсия. Это позволяет получить любое число $m \le k$ дискриминантных компонент, соответствующих *m* минимальным собственным числам пучка {*A*,*B*}.

Элементы вектора *C* можно интерпретировать как веса, отражающие важность каждой компоненты векторов *X*, *Y* при разделении исходных совокупностей. Если используется несколько дискриминантных компонент, это соответствует различным независимым точкам зрения на *X*, *Y.* Иногда дискриминантным компонентам удается приписать некий физический смысл – тогда можно говорить о ***дискриминантных факторах***.

Пример. Для двух данных *k*-мерных выборок $X = [X_1, X_2,\ldots, X_n]$, $Y = [Y_1, Y_2,\ldots, Y_n]$ ($k \ge 2$) найти коэффициенты двух главных дискриминантных факторов по матрицам рассеяния и определить долю информации, объясняемую этими факторами.

Решение.

Процедура вычисления двух дискриминантных факторов по матрицам рассеяния

```
function [c,d,r]=discr0(X,Y);
%X,Y - сравниваемые участки измерений - матрицы (n1xm), (n2xm)
%Проводится линейный дискриминантный анализ по рассеянию
[n1,m]=size(X); [n2,m]=size(Y);
%Вычисление матриц внутригруппового и межгруппового рассеяния
SX=cov(X)*(n1-m); SY=cov(Y)*(n2-m);
S1=SX+SY;
%Вычисление объединенной матрицы рассеяния
Z=[X;Y];
S2=cov(Z)*(n1+n2-m);
%Вариант:  SX=cov(X); SY=cov(Y); S1=SX+SY; Z=[X;Y]; S2=cov(Z)*(n1+n2-m);
[P,Q]=eig(S2,S1);
c=-P(:,m); d=P(:,m-1);  %старшие собственные векторы пучка квадратичных форм
```

Процедура вычисления коэффициентов двух дискриминантных факторов по матрицам внутреннего рассеяния и расстоянию между центрами проекций

```
function [c,d]=discr1(X,Y);
%X,Y - сравниваемые участки измерений - матрицы (n1xm), (n2xm)
%Проводится линейный дискриминантный анализ с регуляризацией
[n1,m]=size(X); [n2,m]=size(Y);
%Вычисление матриц внутригруппового и межгруппового рассеяния
SX=cov(X)*(n1-m);
SY=cov(Y)*(n2-m);
M=n1*mean(X)-n2*mean(Y);
S1=SX+SY;
S2=(M'*M+0.00001*eye(size(S1)));     %регуляризация
[P,Q]=eig(S2,S1);
c=-P(:,m); d=P(:,m-1);  %старшие собственные векторы пучка квадратичных форм
```

Пример. Смоделировать две независимые 3-мерные выборки с заданными параметрами и представить их

- В виде 3-мерных облаков точек ;
- На плоскости 2 главных факторов;
- На плоскости 2 главных дискриминантных факторов.

Решение. Моделирование и отображение двух выборок

```
function modelir;
%Моделирование
n=100;  r12=0.5;  r23=0.3;  r13=0.4;
S0=[1  r12 r13;
r12  1  r23;
r13 r23  1];

[P,Q]=eig(S0);   Q=diag([1 2 0.1]);

ax=[0 -2 0];  SX=P'*Q*P;  X=modl(ax,SX,n);
ay=[0  2 0];  SY=P*Q*P';  Y=modl(ay,SY,n);
%=============Вывод 3-мерных облаков точек================
subplot(2,2,1);
plot3(X(:,1),X(:,2), X(:,3),'*','LineWidth',2); box; grid;
hold on; plot3(Y(:,1),Y(:,2),Y(:,3),'sr','LineWidth',2);
title('Два 3-мерных облака',...
'FontName','Courier New Cyr','FontSize',14,'FontWeight','Bold');
%========Представление на плоскости главных факторов========
[c,d,q]=factor1([X;Y]);
x1=X*c;  y1=X*d;
x2=Y*c;  y2=Y*d;
subplot(2,2,2);
plot(x1,y1,'*','LineWidth',2); grid;
hold on; plot(x2,y2,'sr','LineWidth',2);
title('Главные факторы',...
'FontName','Courier New Cyr','FontSize',14,'FontWeight','Bold');
%========Дискриминация по матрицам рассеяния (док.8.2)=======
[c,d]=discr0(X,Y);  %2 матрицы
x1=X*c;  y1=X*d;
x2=Y*c;  y2=Y*d;
subplot(2,2,3);
plot(x1,y1,'*','LineWidth',2); grid;
hold on; plot(x2,y2,'sr','LineWidth',2);
title('Дискриминация по матрицам рассеяния',...
'FontName','Courier New Cyr','FontSize',14,'FontWeight','Bold');
%====== Дискриминация по расстоянию между центрами (док.8.3)=======
[c,d]=discr1(X,Y);   %2 матрицы и расстояние между центрами проекций
x1=X*c;  y1=X*d;
x2=Y*c;  y2=Y*d;
subplot(2,2,4);
plot(x1,y1,'*','LineWidth',2); grid;
hold on; plot(x2,y2,'sr','LineWidth',2);
title('Дискриминация по расстоянию между центрами',...
'FontName','Courier New Cyr','FontSize',14,'FontWeight','Bold');
```

Моделирование нормальной 3-мерной выборки объема *n* с заданными параметрами

```
function X=modl(a,S,n);
[P,Q]=eig(S);  E0=randn(n,3);   E=E0*Q;  X=E*sqrtm(S);
X(:,1)=a(1)+X(:,1);     X(:,2)=a(2)+X(:,2);     X(:,3)=a(3)+X(:,3);
```

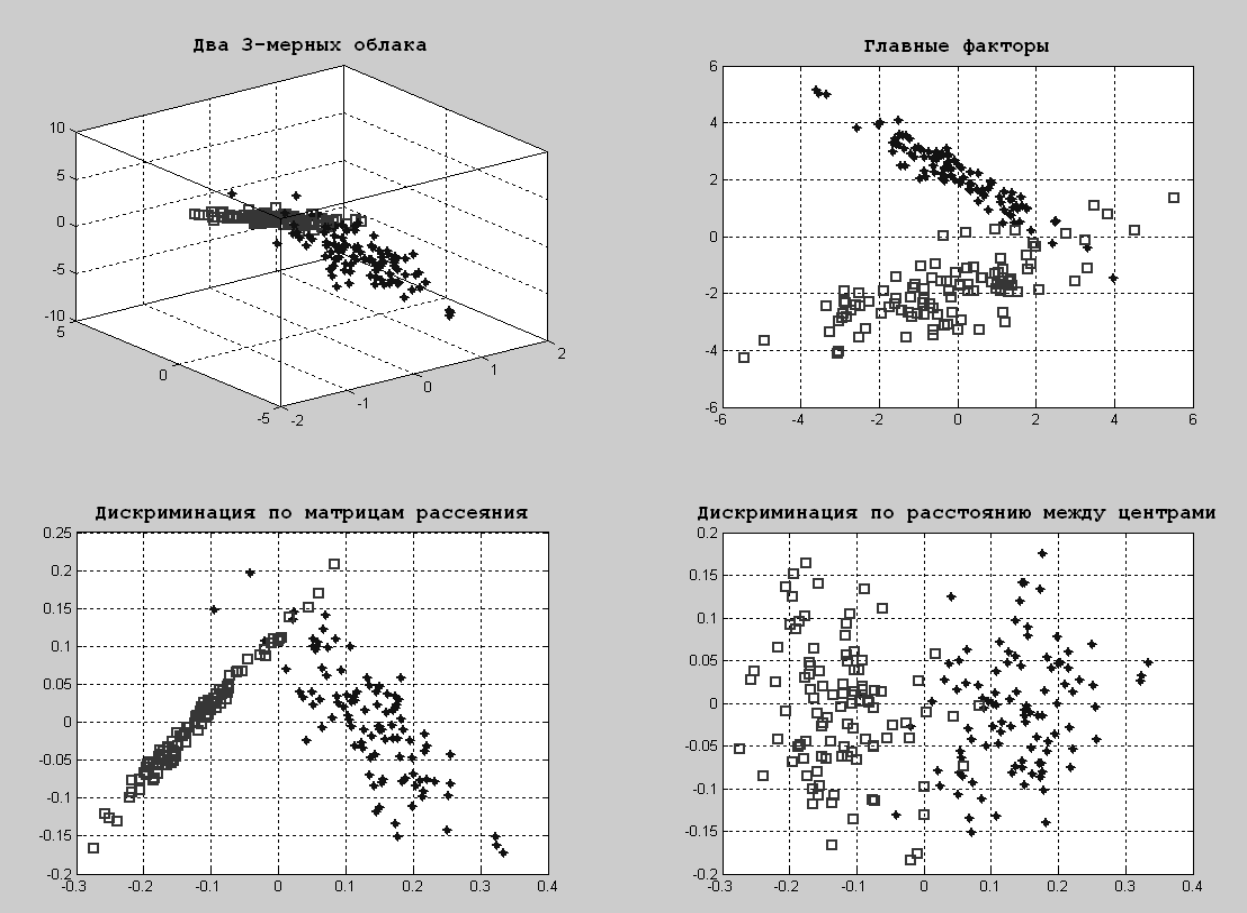


Рис.2. Результат работы программы

ЛЕКЦИЯ 16

КЛАСТЕРИЗАЦИЯ ДАННЫХ

1. Постановка задачи кластеризации

Формальная постановка задачи кластеризации описывается следующим образом.

Пусть задано множество объектов $\{X\}_n$, обладающих k свойствами, и множество имен (меток) кластеров Y. Задана также функция определения схожести объектов $\rho(x_i, x_j)$; $i, j=1,2, \ldots, n$.

Требуется разбить множество X на непересекающиеся подмножества – кластеры – так, чтобы каждый кластер состоял из объектов схожих между собой, а объекты различных кластеров существенно отличались по своим свойствам. При этом каждому объекту из X приписывается имя (номер кластера) из Y.

Например, при рассмотрении множества животных могут быть сформированы кластеры «Большие животные», «Сухопутные животные» и их альтернативы. Впрочем, для альтернатив названных кластеров достаточно затруднительно подобрать осмысленные названия.

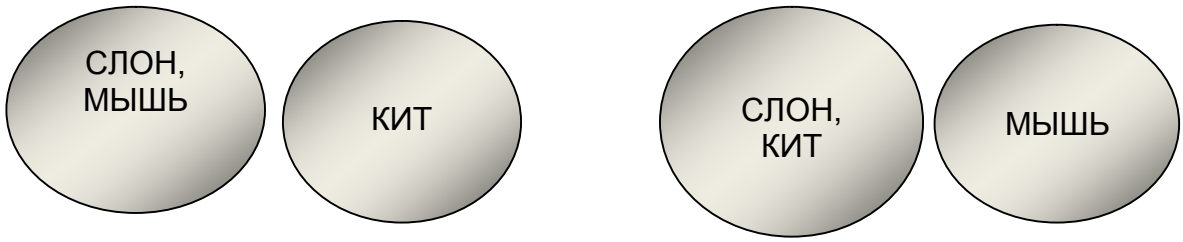


Рис.1. Примеры кластеризации по разным признакам

NB! Напомним, что порядок формирования кластеров определяется перечнем доступных характеристик исследуемого множества объектов.

Кластеризация является важной частью понятия более высокого уровня – кластерного анализа. Второй его этап – интерпретация полученных результатов – необходим для того, чтобы описать полученные результаты в терминах предметной области.

NB! Прародителем кластер-анализа можно считать древнегреческого философа Демокрита. В «Письме учебному соседу» он писал: «Если суть многих вещей тебе непонятна, расположи сходные из них между собой и тебе многое станет яснее».

Кластер-анализ нашел применение в следующих ситуациях:

– понимание данных на основе выявленной кластерной структуры. Сущность данной задачи заключается в группировке объектов по их характеристикам. В данном случае неизвестное свойство – это имя кластера, к которому оно относиться;

– «сжатие» данных путем замены множества объектов кластера его типичным представителем;

– обнаружение закономерностей в наборах данных. В отличие от первой задачи, в этом случае существенно распределение объектов, относящихся к известным классам, по кластерам.

Так, к примеру, решается задача кредитного скоринга – выявление условий (свойств объектов), при которых клиент не склонен к своевременному возвращению кредита;

– уточнение границ классов для распознавания с обучением;

– обнаружение аномалий (Novelty Detection) путем выявления нетипичных объектов, т.е. таких, которые наиболее непохожи на другие объекты кластера.

2. Классификация методов кластеризации

Исчерпывающей и однозначной классификации методов кластеризации еще не разработано. Одной из причин этого является постоянное развитие данного направления анализа, появление новых объектов кластеризации и адекватных им методов обработки данных.

Ниже приводится вариант классификации методов кластеризации объектов в пространстве разнотипных признаков.

По виду искомой кластерной структуры выделяют методы, ориентированные на *разбиения* (рис.2) и *иерархии* (рис.3).

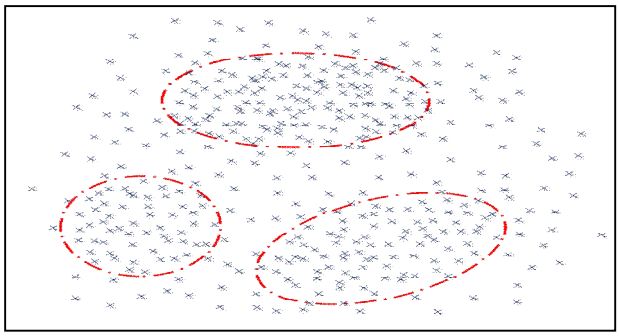


Рис.2. Пример разбиения

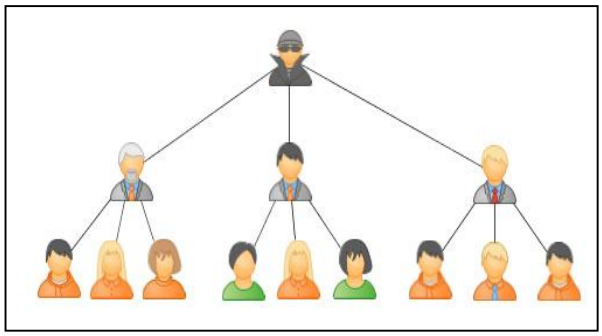


Рис.2. Пример иерархии

Разбиения представляют собой «сгущения» точек, выделенные в пространстве характеристик, т.е. группы наиболее схожих между собой объектов. При этом некоторые точки могут не входить в какой-либо кластер.

Методы, позволяющие строить иерархические структуры (дендрограммы), как правило, итеративны: на каждом шаге формируется один из уровней иерархии. На самом верхнем уровне все объекты исследуемой выборки объединяются в один кластер. Следует отметить, что данная операция не несет какой-либо смысловой нагрузки, но позволяет логически завершить иерархию.

NB! Итеративно разбивая разбиение, полученное на предыдущем шаге, получим иерархическую структуру.

3. Иерархические алгоритмы кластеризации

По способу формирования кластеров в иерархических алгоритмах выделяют *центроидный* и *дисперсионный* методы и *методы связи* (одиночной, полной или средней).

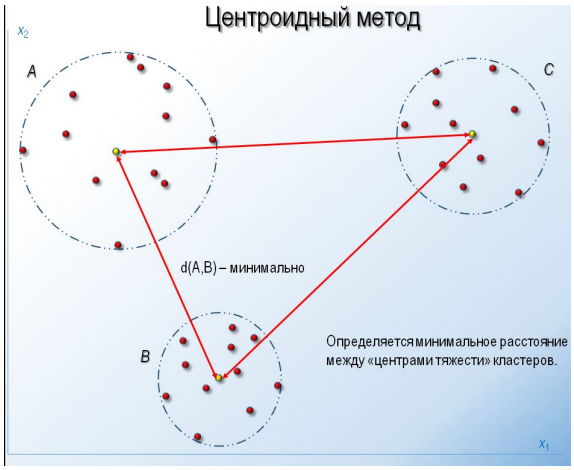


Рис.4. Центроидный метод

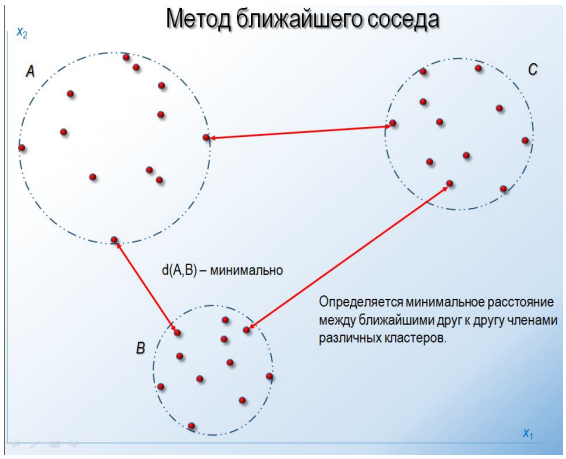


Рис.5. Метод ближайшего соседа

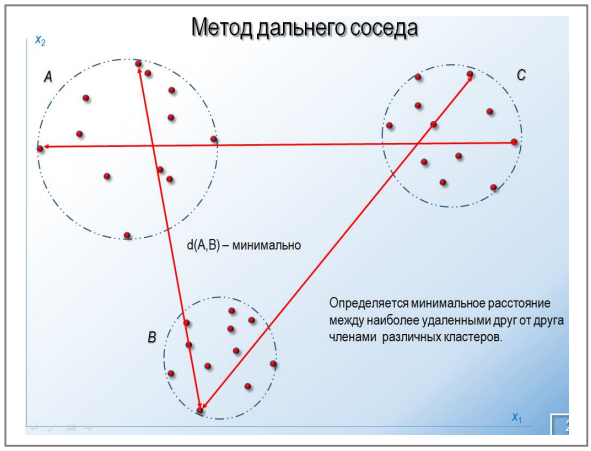


Рис. 6. Метод дальнего соседа

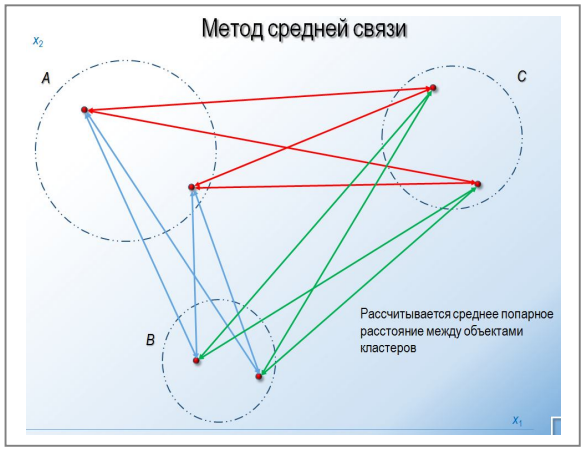


Рис. 7. Метод средней связи

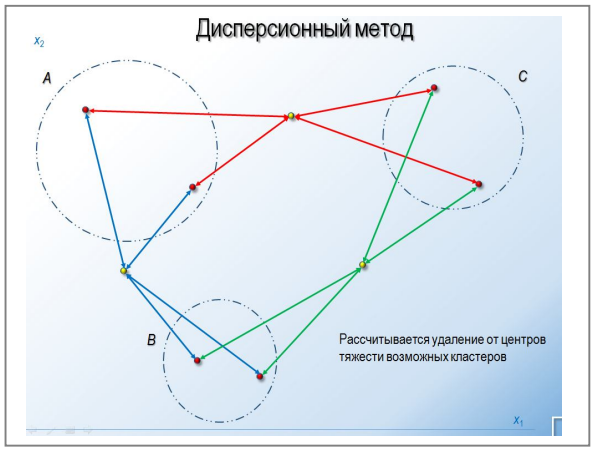


Рис. 8. Дисперсионный метод

Общим правилом является объединение двух кластеров, для которых обеспечивается минимальное значение некоторого показателя:

- для центроидного метода минимальным должно быть расстояние между «центрами тяжести» кластеров;
- для дисперсионного метода минимизируется внутрикластерная дисперсия;
- для метода одиночной связи минимальным должно быть расстояние между ближайшими членами различных кластеров,
- для метода полной связи – расстояние между наиболее удаленными друг от друга объектами, относящимся к различным кластерам;

NB! Метод одиночной связи называют также методом «ближайшего соседа», а полной связи – «дальнего соседа».

- для метода средней связи выбирается минимальное среднее расстояние между объектами различных кластеров.

Один из возможных алгоритмов иерархической кластеризации основан на последовательном объединении объектов в кластеры, а кластеров – во все более крупные кластеры. Эта процедура описывается следующим итеративным алгоритмом.

Каждый объект определяется отдельным кластером;
Пока количество кластеров > 1
Найти пару ближайших объектов (кластеров);
Объединить их в новый кластер;
Конец

Необходимо отметить особенность реализации центроидного метода: в нем при объединении объектов в кластеры происходит физическое замещение объединенных объектов некоторым абстрактным, соответствующем их центру тяжести в пространстве характеристик. Соответственно, на каждой итерации сокращается таблица взаимных расстояний. В других методах (связи, дисперсионных) физического удаления объектов не производится, поэтому таблица расстояний строится только один раз, однако, необходимы дополнительные структуры данных для хранения информации о кластерной структуре (какие объекты в какие кластеры входят).

Следует также сказать, что кроме рассмотренного выше алгоритма объединения объектов или кластеров (*агломеративного*), существуют и *дивизимные алгоритмы* кластеризации, в которых исследуемое множество объектов объединено в единый кластер, а далее проводится его декомпозиция. При этом результаты агломеративной и дивизимной кластеризации при одинаковом правиле объединения объектов и одной и той же выборке данных могут не совпадать.

Важным для анализа дендрограмм является расстояние между кластерами. Напомним, что чем меньше расстояние, тем больше сходство, поэтому наиболее «плотные» кластеры образуются при малых расстояниях.

При этом возможно существование объектов, представляющих собой аномалии, поскольку значительно удалены как друг от друга, так и от других объектов.

Несомненными *достоинствами алгоритмов иерархической кластеризации* полагают:

- возможность выбора необходимого количества кластеров на основе анализа дендрограммы;

- независимость результата от начальных условий (при заданном методе объединения кластеров анализ одной и той же выборки данных дает одни и те же результаты).

Недостатком же является ограниченное количество объектов анализа: дендрограмма с количеством объектов более сотни воспринимается с трудом. Именно вследствие этого недостатка в эпоху Больших данных (Big Data) развиваются неиерархические алгоритмы.

4. Неиерархические алгоритмы кластеризации

Неиерархические алгоритмы кластеризации ориентированы на построение разбиений. Ввиду значительного разнообразия подходов рассмотрим только некоторые из алгоритмов.

Достаточно простым и понятным является *алгоритм кратчайшего незамкнутого пути*, основанный на удалении самых длинных ребер в кратчайшем незамкнутом пути, построенном на полносвязном графе. Построение кратчайшего незамкнутого пути можно представить следующим алгоритмом:

```
// Дано: матрица смежности полносвязного графа
// Найти: разбиение графа на n кластеров
// Построение кратчайшего незамкнутого пути:
Начало:
Найти пару элементов с минимальным расстоянием;
Соединить их;
Повторять пока есть свободные элементы
 Найти элемент, ближайший к одной из конечных точек связанного пути;
    Соединить их;
// Конец построения КНП
Повторять n-1 раз
    Найти самое длинное ребро КНП;
    Удалить его;
Конец. // Получили n кластеров
```

Другой распространенный алгоритм носит название k-средних (k-means). Он обеспечивает построение заданного числа (*k*) кластеров с минимальным квадратичным отклонением их членов от центров этих кластеров.

Динамика алгоритма k-means показана на рис. 9.

Достоинствами алгоритма k-means полагают высокую скорость работы и простоту использования.

Недостатков у данного алгоритма гораздо больше:

- чувствительность к выбору начальных условий (первичных центров кластеров);

- чувствительность к выбросам данных (аномальным значениям);

- существенное замедление по мере роста данных

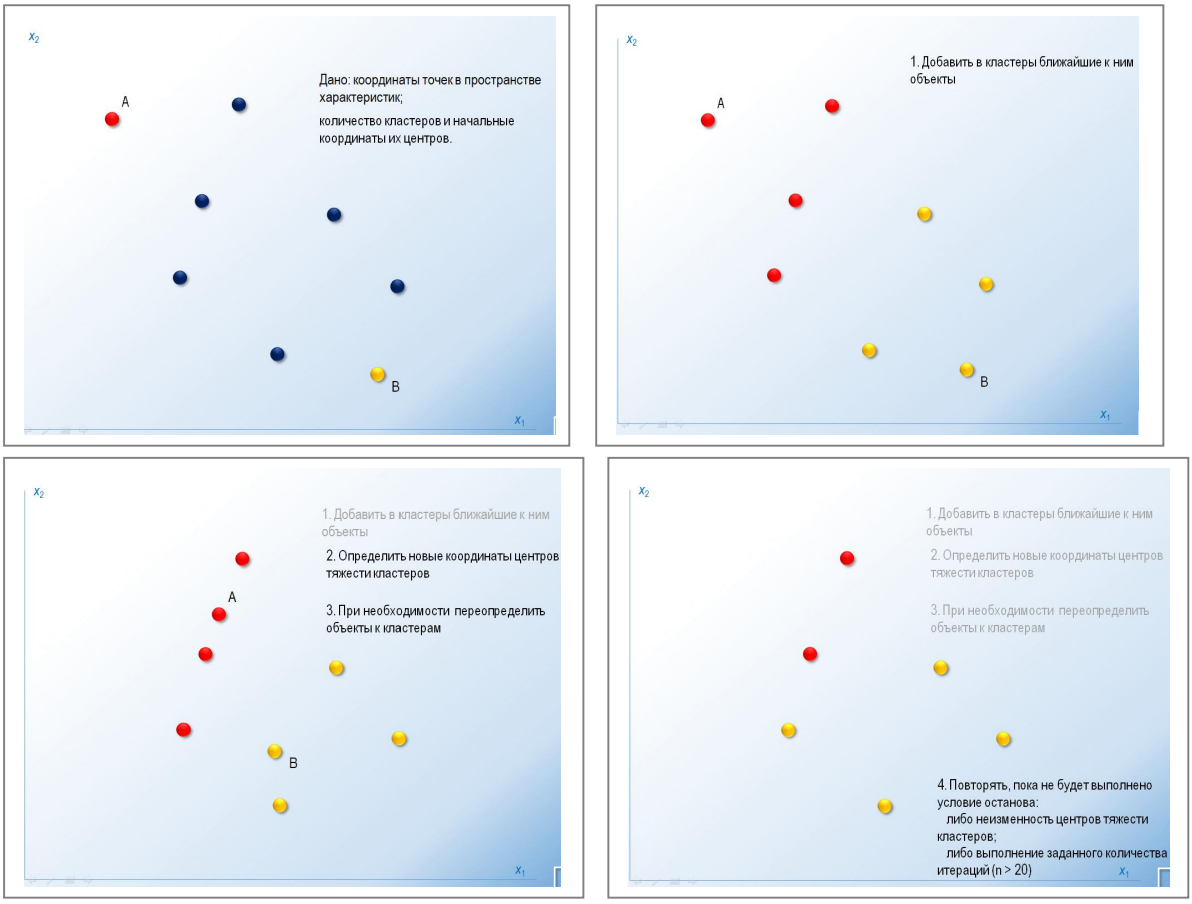


Рис. 9. Динамика алгоритма k-means

Тем не менее, этот алгоритм используется как предварительный фильтр для нейросети, обеспечивающий обработку изображений.

5. Качество кластеризации

Основной проблемой, возникающей при оценивании качества кластеризации является сложность математической формулировки задачи *улучшения понимания структуры данных*. Эта проблема обуславливает наличие широкого спектра процедур формирования оценки качества кластеризации.

Интуитивно понятно, что качество кластеризации тем выше, чем меньше внутрикластерное расстояние и, соответственно, больше межкластерное. Что Вам напоминает это выражение? Как вариант, для сравнения двух кластерных структур критерием качества может служить минимальная дисперсия объектов в кластере или внутригрупповая функция квадратов отклонения:

$$\theta = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

где x_i – значение i-й характеристики объекта.

NB! Эти способы применимы только при известном количестве кластеров!

Еще один возможный критерий – это *стабильность кластерной структуры* при добавлении в выборку новых объектов. При незначительных отклонениях значений характеристик новых объектов от уже имеющихся кластерная структура должна либо сохраняться, либо меняться незначительно.

Стабильность кластерной структуры может быть оценена и другим способом: применением других алгоритмов и способов кластеризации. Схожесть кластерных структур, полученных при использовании различных подходов, указывает на «объективность» кластеризации.

Заканчивая обзор методов кластеризации, следует, хотя бы, упомянуть те, которые получили достаточно широкое распространение в науке об обработке данных, но не рассматриваются в лекции в силу невозможности объять необъятное. К ним относятся: алгоритмы нечеткой кластеризации (fuzzy c-means), нейросетевые технологии – самоорганизующиеся карты Кохонена, (Self-Organized Maps) и т.п.

Вопросы:

1. Приведите формализованную постановку задачи кластеризации.
2. Где используется кластер-анализ?
3. Приведите общее правило объединения двух кластеров.
4. Что называется разбиением? Иерархией?
5. В чем состоит метод «ближайшего соседа»? «Дальнего соседа»?
6. Что называется методом средней связи?
7. Назовите особенности реализации центроидного метода.
8. Чем отличаются агломеративные и дивизимные алгоритмы кластеризации?
9. Назовите достоинства алгоритмов иерархической кластеризации.
10. В чем состоит алгоритм кратчайшего незамкнутого пути?
11. В чем состоит алгоритм k-средних? Его достоинства и недостатки?
12. Что называется внутрикластерным расстоянием? Межкластерным расстоянием?
13. Как определяется стабильность кластерной структуры?

Литература:

. Классификация и снижение размерности/Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Под ред. С.А.Айвазяна.- М.: Финансы и статистика, 1989.- 607с.

. Вентцель Е.С. Теория вероятностей. - М.: Физматгиз, 1962. - 564с.

. Прикладная статистика. Исследование зависимостей/// Айвазян С.А., Енюков И.С., Мешалкин Л.Д. /Под ред. С.А.Айвазяна.- М.: Финансы и статистика, 1989.- 607с.

. Кендалл М., Стьюарт А. Статистические выводы и связи// Пер. с англ. под ред. А.Н.Колмогорова.- М.: Наука, 1973.- 900с.

ЛЕКЦИЯ 17

СОКРАЩЕНИЕ РАЗМЕРНОСТИ И ВИЗУАЛИЗАЦИЯ

Вопросы:

1. Главные компоненты и факторный анализ
2. Нелинейные главные компоненты (проектирование с контрастированием)

1. Главные компоненты и факторный анализ

Пусть случайный вектор $X = (x_1, \ldots, x_k)^T \in N_k(a, \Sigma)$. Первой главной компонентой вектора X называется линейная комбинация

$$z_1 = \sum_{i=1}^{k} c_{1i} x_i = C_1^T X; \quad C_1 = (c_{11},\ldots,c_{1k})^T, \quad \|C_1\|^2 = \sum_{i=1}^{k} c_{1i}^2 = 1,$$

где коэффициенты c_{1i} выбираются так, чтобы величина z_1 имела наибольшую дисперсию среди всех нормированных линейных комбинаций компонент x_i.

Дисперсия случайной величины $z = C^T X$

$$D(C^T X) = M[(C^T X)(C^T X)^T] = M(C^T X X^T C) = C^T [M(XX^T)]C = C^T \Sigma C,$$

поэтому вектор C_1 является решением оптимизационной задачи $C^T \Sigma C \to \max$ с ограничением $\|C\|^2 = C^T C = 1$.

Составим для этой задачи функцию Лагранжа

$$L(C, \lambda) = C^T \Sigma C - \lambda (C^T C - 1),$$

затем приравняем нулю ее производные по векторному аргументу C и скалярному аргументу λ :

$$\begin{cases} \dfrac{\partial L}{\partial C} = 2\Sigma C - 2\lambda C = 0 \\ \dfrac{\partial L}{\partial \lambda} = -(C^T C - 1) = 0 \end{cases}.$$

Первое уравнение приводится к виду $(\Sigma - \lambda I)C = 0$: точками, подозрительными на экстремум, являются пары собственных чисел и соответствующих собственных векторов матрицы Σ.

Второе уравнение показывает, что собственные векторы должны иметь единичную длину.

Подставим найденные варианты значений вектора C в исходную целевую функцию:

$$C^T \Sigma C = C^T \cdot \lambda C = \lambda \cdot (C^T C) = \lambda.$$

Таким образом, максимальное значение дисперсии z, равное $\lambda_{max}(\Sigma)$, достигается, если вектор C – нормированный собственный вектор ковариационной матрицы Σ, соответствующий ее максимальному собственному числу $\lambda_{max}(\Sigma)$.

Геометрически это означает, что вектор C_1 параллелен наибольшей оси эллипсоида рассеяния вектора X. Поскольку суммарная дисперсия всех компонент вектора X

$$\sum_{i=1}^{k} Dx_i = \sum_{i=1}^{k} \sigma_{ii}^2 = tr(\Sigma) = \sum_{i=1}^{k} \lambda_i,$$

говорят, что доля суммарной дисперсии, объясняемая первой главной компонентой z_1, равна

$$\frac{\lambda_1}{\lambda_1 + \ldots + \lambda_k} = \frac{\lambda_1}{tr(\Sigma)}.$$

Аналогично, с использованием второго по величине собственного числа λ_2 и соответствующего собственного вектора C_2, ортогонального, как известно, C_1, определяется вторая главная компонента и т.д. Векторы C_1, C_2 и т.д. можно получать также непосредственно из решения оптимизационных задач:

$$C_1 = \arg\max \frac{D(C^T X)}{\sum_{i=1}^{k} Dx_i}, \quad \|C\| = 1;$$

$$C_2 = \arg\max \frac{D(C^T X)}{\sum_{i=1}^{k} Dx_i}, \quad \|C_1\| = 1, \quad C \perp C_1$$

и т.д. Геометрически переход от вектора X к его первым двум главным компонентам означает его проектирование на плоскость, параллельную главным осям эллипсоида рассеяния. На практике матрица Σ обычно неизвестна, и ее заменяют матрицей $\hat{\Sigma}$, полученной на основе обучающей выборки $X_1, \ldots, X_n$.

Пример 1. Определение линейной главной компоненты для двумерного вектора.

Пусть

$$X = (x, y)^T \in N_2(0, \Sigma); \qquad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 2 \end{bmatrix}.$$

Прежде всего, находятся собственные числа матрицы Σ:

$$\det\begin{bmatrix} 1-\lambda & -0.5 \\ -0.5 & 2-\lambda \end{bmatrix} = 0: \quad \lambda_1 = 2.21, \quad \lambda_2 = 0.79.$$

Далее определяется собственный вектор $C = (c_1, c_2)^T$, соответствующий наибольшему собственному числу $\lambda_1 = 2.21$, из системы уравнений

$$\begin{cases} -1.21c_1 - 0.5c_2 = 0 \\ c_1^2 + c_2^2 = 1 \end{cases}: \quad C = \begin{bmatrix} 0.38 \\ -0.92 \end{bmatrix}.$$

Этот вектор C задает направление главной оси x_1. Ось y_1 ей перпендикулярна; ее направление можно также определить как собственный вектор, соответствующий второму собственному числу λ_2. Ортогональность этих осей обеспечивается симметрией матрицы Σ.

На рис. 1 приведен эллипс рассеяния для вектора X. Направление его главной оси задается найденным вектором C. В осях x_1, y_1 этот эллипс описывается уравнением

$$\frac{x_1^2}{(\sqrt{\lambda_1})^2} + \frac{x_2^2}{(\sqrt{\lambda_2})^2} = 1.$$

Доля дисперсии, объясняемой первой главной компонентой, $\lambda_1/(\lambda_1 + \lambda_2) = 74\%$. Таким образом, размерность вектора уменьшается в два раза, а эффективность представления - только на 26%.

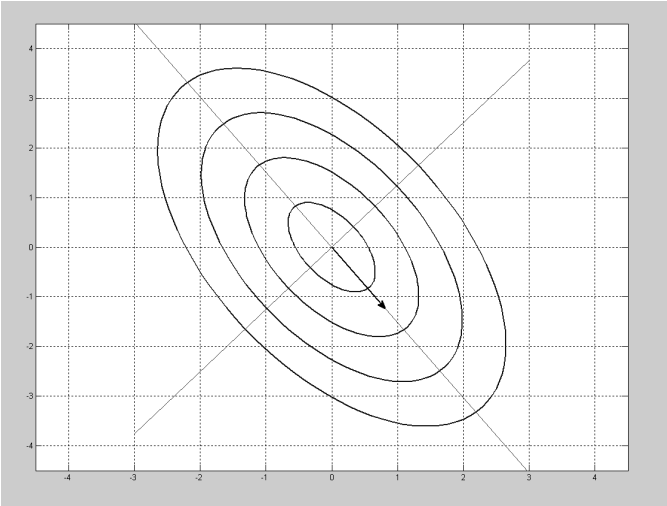


Рис. 1. Семейство эллипсов рассеяния в примере 1

Данную технику, как правило, применяют при анализе *k*-мерной совокупности однотипных измерений – *k*-мерного облака точек $X_1,\ldots, X_n$, которые трактуют как независимые измерения случайного вектора *X.* Для параметров *a,* $\sum$ используют их выборочные оценки.

Если компоненты вектора *X* имеют различную физическую природу и измерены с помощью качественно различных технических средств, то результат не имеет, как правило, ясного физического смысла и существенно зависит от выбора масштаба. В этих случаях вместо ковариационной матрицы $\sum$ или ее оценки используют корреляционную матрицу с элементами

$$r_{ij} = \sigma_{ij} (\sigma_{ii} \sigma_{jj})^{-1/2} ,$$

которые являются безразмерными величинами (диагональные элементы $r_{ii} = 1$).

Главные компоненты, построенные по корреляционной матрице – безразмерные случайные величины – называют *главными факторами*, объясняющими рассеяние компонент вектора *X*.

Традиционный факторный анализ использует несколько иную технику вычислений коэффициентов c_{ij} – факторных нагрузок, основанную на методе наименьших квадратов. Его результаты обычно очень близки к результатам анализа главных компонент по корреляционной матрице, однако в факторный анализ органически входит специальный метод вращений, призванный облегчить содержательную интерпретацию получающихся факторов.

Пример 2. Для данной *k*-мерной выборки $X = [X_1, X_2,\ldots, X_n]$ ($k \geq 2$) найти коэффициенты двух главных факторов и определить долю информации, объясняемую этими факторами.

Процедура вычисления коэффициентов двух главных факторов

```
function [c,d,r]=factor1(X);
%Анализ главных компонент
[n,m]=size(X);
[P,Q]=eig(corrcoef(X));  %главные компоненты по корреляционной матрице
c1=P(:,m); c2=P(:,m-1);
q=diag(Q);
r=(q(m)+q(m-1))/sum(q); % контроль объясняемой дисперсии
```

Есть два пути для выявления физического смысла найденных факторов. Во-первых, можно проанализировать веса c_{ij}. Во-вторых, можно просто расположить исходные объекты, описываемые измерениями X, в порядке возрастания i-го фактора и попытаться понять, какому физическому свойству соответствует такое упорядочение.

Линейные главные компоненты обладают целым рядом других оптимальных свойств, которые в некоторых случаях можно использовать даже в качестве их начального определения. Пусть $X_1, \dots, X_n$, $X_i = (x_{i1}, \dots, x_{ik})^T$ - независимая выборка из распределения $N_k(a, \Sigma)$, а $z_1, \dots, z_n$ - их образы при линейном проектировании в некоторое q - мерное пространство ($q < k$):

$$X_i \to Z_i = (z_{1i}, \dots, z_{iq})^T, \quad z_{il} = \sum_{j=1}^{k} c_{lj} x_{ij} = C^T X_i.$$

Пусть d_{ij} - расстояние между X_i и X_j в евклидовом пространстве R^k, ρ_{ij} - расстояние между их образами в R^q. В качестве меры искажения матрицы попарных расстояний используется величина

$$\Delta^2 = \sum_{i,j=1}^{n} (d_{ij} - r_{ij})^2. \qquad (1)$$

Имеет место следующее важное свойство: Δ_q^2 минимально, если в качестве Z_i взяты первые q главных компонент векторов $X_1, \dots, X_n$. На этом свойстве, в частности, основаны различные нелинейные аналоги метода главных компонент, которые получаются, если рассматривать меры искажения, отличные от (1).

Пример 3. Несколько лет назад в одном из российских НИИ решалась следующая задача. Имелось около 300 образцов смазочных материалов, отечественных и импортных. Каждый образец анализировался по 9 параметрам. Таким образом, исходная информация представляла собой облако точек в 9-мерном пространстве. При проектировании этого облака на плоскость двух первых главных факторов величина

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{9} \lambda_i}$$

оказалась равной 0.64, т.е. потери информации при проектировании составили около 36% . Анализ выявленных факторов позволил выделить 14 сгустков (кластеров), в каждый из которых вошли материалы с достаточно близкими свойствами, что дало основания для рекомендаций по замене импортных материалов их отечественными аналогами. 32 точки при этом оказались изолированными - это материалы, не имеющие аналогов и не допускающие замены.

2. Нелинейные главные компоненты (Проектирование с контрастированием)

Пусть в k-мерном пространстве имеется облако точек $X_1, \dots, X_n$, d_{ij} - расстояние между точками X_i и X_j. Пусть эти точки каким-то образом проектируются в пространство R^q, $q<k$: $X_i \to Z_i \in R^q$, ρ_{ij} - расстояние между Z_i и Z_j. В качестве меры искажения геометрической структуры данных при переходе от $\{X_i\}$ к $\{Z_j\}$ вводится функция

$$Q_\alpha(z_1,...,z_n) = \frac{\sum_{i>j}(d_{ij} - \rho_{ij})^2 d_{ij}^\alpha}{\sum_{i>j} d_{ij}^\alpha}$$

Образы z_1 , ... , z_n ищутся из условия Q_α (z_1 , ... , z_n) = min . При $\alpha < 0$ метод более чувствителен к искажению малых расстояний , при $\alpha > 0$ - больших. Рекомендуется использовать значение $\alpha = -1$. Такое проектирование приводит к контрастному выделению неоднородностей (сгустков) в множестве $\{X_i\}$, а также выявлению среди них аномальных значений (выбросов). Задача решается итерационным градиентным методом:

$$z_i^{(k+1)} = z_i^{(k)} + \left[2\sum_{j=1}^{n} d_{ij}^{\alpha} \right]^{-1} \frac{\partial Q_\alpha}{\partial z_i}$$

В качестве начального приближения используются линейные главные компоненты. Можно рассматривать и другие конструкции мер искажения.

Подобные задачи называют задачами редукции размерности или визуализации. Результаты рационального снижения размерности часто интерпретируют как нахождение основополагающих скрытых характеристик (факторов) имеющегося массива данных. В наибольшей степени эта тенденция нашла свое выражение в концепции факторного анализа. Действительно, можно доказать, что если такие скрытые переменные в данных действительно есть и они независимы между собой, то в результате осуществления операций по рациональному снижению размерности они, как правило, обнаруживаются.

Вопросы для самопроверки:

1. Что называется главной компонентой?
2. Чему равна дисперсия главной компоненты?
3. Что называется функцией Лагранжа?
4. Что означает геометрически переход от вектора X к его первым двум главным компонентам?
5. Как определяется физический смысл главных компонент?
6. Чем отличается факторный анализ от метода главных компонент?
7. Какими оптимальными свойствами обладают линейные главные компоненты?
8. Какая величина используется в качестве меры искажения матрицы попарных расстояний?
9. Каким образом осуществляется проектирование с контрастированием?

III. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ КОМПЬЮТЕРНОЙ МАТЕМАТИКИ

ЛЕКЦИЯ 18
ВВЕДЕНИЕ В ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

ИНС предназначены для моделирования деятельности мозга – сознания. ИНС чрезвычайно разнообразны по своим конфигурациям. История развития этого вопроса – от модели нейрона до развивающихся многослойных сетей приведены в настоящей лекции.

1. Введение. Исторический аспект

Людей всегда интересовало их собственное мышление. Эти попытки самопознания, размышления мозга о себе самом является, возможно, одной из отличительных особенностей человека.

Имеется множество размышлений о природе мышления, простирающихся от анатомических до духовных. Обсуждение этого вопроса, протекавшее в горячих спорах философов и теологов с физиологами, анатомами и кибернетиками, принесло мало пользы, сам предмет исследования оказался слишком труден для изучения. Те, кто опирался на самоанализ и эмпирические размышления, пришли к выводам, не отвечающим уровню строгости физических наук. Экспериментаторы же нашли, что мозг труден для наблюдения и ставит в тупик своей организацией. Короче говоря, мощные методы научного исследования, изменившие наш взгляд на физическую реальность, оказались бессильными в понимании самого человека.

Нейробиологи и нейроанатомы достигли значительного прогресса. Усердно изучая структуру и функции нервной системы человека, они многое поняли в «электропроводке» мозга, но мало узнали о его функционировании. В процессе накопления ими знаний выяснилось, что мозг имеет ошеломляющую сложность. Десятки миллиардов нейронов, каждый из которых соединен с сотнями или тысячами других, образуют систему, далеко превосходящую наши самые смелые мечты о суперкомпьютерах. Тем не менее, мозг постепенно выдает свои секреты в процессе одного из самых напряженных и честолюбивых исследований в истории человечества.

Лучшее понимание функционирования нейрона и картины его связей позволило исследователям создать математические модели для проверки своих теорий. Эксперименты теперь могут проводиться на компьютерах без привлечения человека или животных, что решает многие практические и морально-этические проблемы.

В первых же работах выяснилось, что эти модели не только повторяют функции мозга, но и способны выполнять функции, имеющие свою собственную ценность. Поэтому возникли и остаются в настоящее время две взаимно обогащающие друг друга цели *нейронного моделирования*: первая – понять функционирование нервной системы человека на уровне физиологии и психологии и вторая – создать вычислительные системы, выполняющие функции, сходные с функциями мозга.

Параллельно с прогрессом в нейроанатомии и нейрофизиологии психологами были созданы модели человеческого обучения. Одной из таких моделей, оказавшейся наиболее плодотворной, была модель Д. Хэбба, который в 1949г. предложил закон обучения, явившийся стартовой точкой для алгоритмов обуче-

ния *искусственных нейронных сетей (ИНС)*. Дополненный сегодня множеством других методов он продемонстрировал ученым того времени, как сеть нейронов может обучаться.

В пятидесятые и шестидесятые годы группа исследователей, объединив эти биологические и физиологические подходы, создала первые ИНС. Выполненные первоначально как электронные сети, они были позднее перенесены в более гибкую среду компьютерного моделирования, сохранившуюся и в настоящее время.

Первые успехи вызвали взрыв активности и оптимизма. Минский, Розенблатт, Уидроу и другие разработали сети, состоящие из одного слоя искусственных нейронов, называемые персептронами. Такие сети были использованы для такого широкого класса задач, как предсказание погоды, анализ электрокардиограмм и искусственное зрение. В течение некоторого времени казалось, что ключ к интеллекту найден и воспроизведение человеческого мозга является лишь вопросом конструирования достаточно большой сети.

Но эта иллюзия скоро рассеялась. Сети не могли решать задачи, внешне весьма сходные с теми, которые успешно решал мозг. Минский, используя точные математические методы, строго доказал ряд теорем, относящихся к функционированию сетей. Его исследования привели к написанию книги [4], в которой он вместе с Пайпертом доказал, что используемые в то время однослойные сети теоретически неспособны решить многие простые задачи, в том числе реализовать функцию «Исключающее ИЛИ». Минский также не был оптимистичен относительно потенциально возможного здесь прогресса.

Персептрон показал себя заслуживающим изучения, несмотря на жесткие ограничения (и даже благодаря им). У него много привлекательных свойств: линейность, обучаемость, простота модели параллельных вычислений. Нет оснований полагать, что эти достоинства сохраняться при переходе к многослойным системам. Возможно, будет открыта какая-то мощная теорема о сходимости или найдена глубокая причина неудач дать интересную «теорему обучения» для многослойных машин ([4], с.231-232).

Блеск и строгость аргументации Минского, а также его престиж породили огромное доверие к его выводам. Разочарованные исследователи оставили поле исследований ради более обещающих областей, а правительства перераспределили свои субсидии, и ИНС были забыты почти на два десятилетия. Тем не менее несколько наиболее настойчивых ученых, таких как Кохонен, Гроссберг, Андерсон продолжили исследования. Наряду с плохим финансированием и недостаточной оценкой ряд исследователей испытывал затруднения с публикациями. Поэтому исследования, опубликованные в семидесятые и начале восьмидесятых годов, разбросаны в массе различных журналов, некоторые из которых малоизвестны. Постепенно появился теоретический фундамент, на основе которого сегодня конструируются наиболее мощные многослойные сети. Оценка Минского оказалась излишне пессимистичной, многие из поставленных в его книге задач решаются сейчас сетями с помощью стандартных процедур.

За последние несколько лет теория ИНН стала широко применяться в прикладных областях, появились новые корпорации, занимающиеся коммерческим использованием этой технологии. Нарастание научной активности носило взрывной характер.

Урок, который можно извлечь из этой истории, выражается законом Кларка (выдвинутым писателем-фантастом Артуром Кларком). В нем утверждается, что, если крупный уважаемый ученый говорит, что нечто может быть выполнено, то он

(или она) почти всегда прав. Если же ученый говорит, что это не может быть выполнено, то он (или она) почти всегда не прав.

История науки является летописью ошибок и частичных истин. То, что сегодня не подвергается сомнениям, завтра отвергается. Некритическое восприятие «фактов» независимо от их источника может парализовать научный поиск. С одной стороны, блестящая научная работа Минского задержала развитие искусственных нейронных сетей. Нет сомнений, однако, в том, что область пострадала вследствие необоснованного оптимизма и отсутствия достаточной теоретической базы. И возможно, что шок, вызванный книгой «Персептроны», обеспечил необходимый для созревания этой научной области период.

2. Биологический прототип

Развитие ИНС вдохновляется биологией. Рассматривая сетевые конфигурации и алгоритмы, исследователи мыслят их в терминах организации мозговой деятельности. Но на этом аналогия может и закончиться. Наши знания о работе мозга столь ограничены, что мало бы нашлось руководящих ориентиров для тех, кто стал бы ему подражать. Поэтому разработчикам сетей приходится выходить за пределы современных биологических знаний в поисках структур, способных выполнять полезные функции. Во многих случаях это приводит к необходимости отказа от биологического правдоподобия, мозг становится просто метафорой, и создаются сети, невозможные в живой материи или требующие неправдоподобно больших допущений об анатомии и функционировании мозга.

Несмотря на то, что связь с биологией слаба и зачастую несущественна, ИНС продолжают сравниваться с мозгом. Их функционирование часто напоминает человеческое познание, поэтому трудно избежать этой аналогии. К сожалению, такие сравнения неплодотворны и создают неоправданные ожидания, неизбежно ведущие к разочарованию. Исследовательский энтузиазм, основанный на ложных надеждах, может испариться, столкнувшись с суровой действительностью, как это уже однажды было в шестидесятые годы, и многообещающая область снова придет в упадок, если не будет соблюдаться необходимая сдержанность.

Несмотря на сделанные предупреждения, полезно все же знать кое-что о нервной системе млекопитающих, так как она успешно решает задачи, к выполнению которых лишь стремятся ИНН.

Последующее обсуждение весьма кратко. Приложение А содержит более обширное (но ни в коем случае не полное) рассмотрение нервной системы млекопитающих для тех, кто хочет узнать больше об этом предмете.

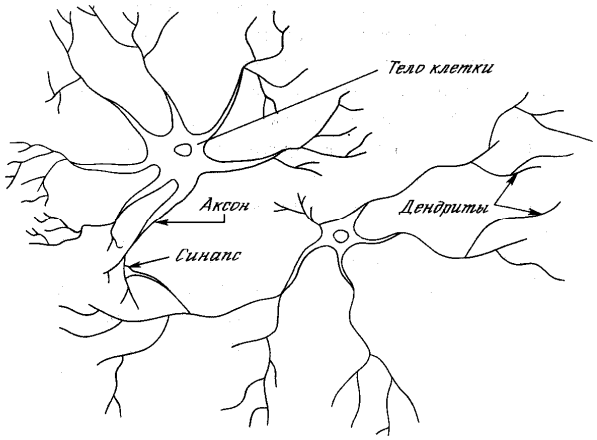


Рис. 1. Биологический нейрон

Нервная система человека, построенная из элементов, называемых нейронами, имеет ошеломляющую сложность. Около 10^{11} нейронов участвуют в примерно 10^{15} передающих связях, имеющих длину метр и более. Каждый нейрон обладает многими качествами, общими с другими элементами тела, но его уникальной способностью является прием, обработка и передача электрохимических сигналов по нервным путям, которые образуют коммуникационную систему мозга.

На рис. 1 показана структура пары типичных биологических нейронов. *Дендриты* идут от тела нервной клетки к другим нейронам, где они принимают сигналы в точках соединения, называемых *синапсами*. Принятые синапсом входные сигналы подводятся к телу нейрона. Здесь они суммируются, причем одни входы стремятся возбудить нейрон, другие – воспрепятствовать его возбуждению. Когда суммарное возбуждение в теле нейрона превышает некоторый порог, нейрон возбуждается, посылая по *аксону* сигнал другим нейронам. У этой основной функциональной схемы много усложнений и исключений, тем не менее большинство ИНС моделируют лишь эти простые свойства.

2. Искусственный нейрон

Искусственный нейрон имитирует в первом приближении свойства биологического нейрона. На вход искусственного нейрона поступает некоторое множество сигналов, каждый из которых является выходом другого нейрона. Каждый вход умножается на соответствующий вес, аналогичный *синаптической силе*, и все произведения суммируются, определяя *уровень активации* нейрона.

На рис. 2 представлена модель, реализующая эту идею.

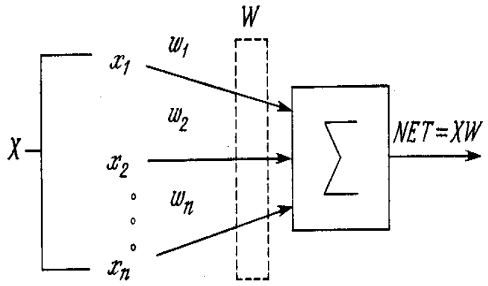


Рис. 2. Искусственный нейрон

Хотя сетевые парадигмы весьма разнообразны, в основе почти всех их лежит эта конфигурация. Здесь множество входных сигналов, обозначенных x_1, x_2,…, x_n, поступает на искусственный нейрон. Эти входные сигналы, в совокупности обозначаемые вектором X, соответствуют сигналам, приходящим в синапсы биологического нейрона. Каждый сигнал умножается на соответствующий вес w_1, w_2, …, w_n, и поступает на суммирующий блок, обозначенный Σ.

Каждый вес соответствует «силе» одной биологической синаптической связи. (Множество весов в совокупности обозначается вектором W.) Суммирующий блок, соответствующий телу биологического элемента, складывает взвешенные входы алгебраически, создавая *выход, который мы будем называть NET*. В векторных обозначениях это может быть компактно записано следующим образом:

NET = XW.

Активационные функции. Сигнал NET далее, как правило, преобразуется *активационной функцией F* и дает выходной нейронный сигнал *OUT*. Активационная функция может быть обычной линейной функцией OUT = *K*(NET), где *K* – постоянная, пороговой функции

OUT = 1, если NET > T,

OUT = 0 в остальных случаях,

где *T* – некоторая *постоянная пороговая величина*, или же функцией, более точно моделирующей нелинейную передаточную характеристику биологического нейрона и представляющей нейронной сети большие возможности.

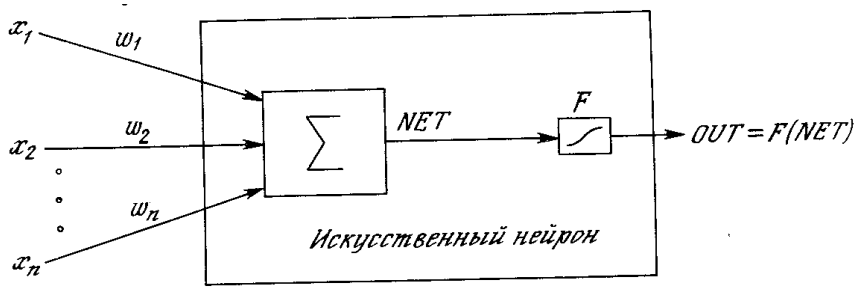


Рис. 3. Искусственный нейрон с активационной функцией

На рис. 3 блок, обозначенный *F,* принимает сигнал NET и выдает сигнал OUT. Если блок *F* сужает диапазон изменения величины NET так, что при любых значениях NET значения OUT принадлежат некоторому конечному интервалу, то *F* называется *«сжимающей» функцией.*

В качестве «сжимающей» функции часто используется *логистическая или «сигмоидальная» (S-образная)* функция, показанная на рис. 4. Эта функция математически выражается как $F(x) = 1/(1 + e^{-x})$. Таким образом,

$$\mathrm{OUT} = \frac{1}{1 + \mathrm{e}^{-\mathrm{NET}}} = \mathrm{F(NET)}.$$

Рис. 4. Сигмоидальная логистическая функция

По аналогии с электронными системами активационную функцию можно считать нелинейной усилительной характеристикой искусственного нейрона.

Коэффициент усиления вычисляется как отношение приращения величины OUT к вызвавшему его небольшому приращению величины NET. Он выражается наклоном кривой при определенном уровне возбуждения и изменяется от малых значений при больших отрицательных возбуждениях (кривая почти горизонтальна) до максимального значения при нулевом возбуждении и снова уменьшается, когда возбуждение становится большим положительным.

Гроссберг (1973) обнаружил, что подобная нелинейная характеристика решает поставленную им *дилемму шумового насыщения*. Каким образом одна и та же сеть может обрабатывать как слабые, так и сильные сигналы? Слабые сигналы нуждаются в большом сетевом усилении, чтобы дать пригодный к использованию выходной сигнал. Однако усилительные каскады с большими коэффициентами усиления могут привести к насыщению выхода шумами усилителей (случайными флуктуациями), которые присутствуют в любой физически реализованной сети. Сильные входные сигналы в свою очередь также будут приводить к *насыщению усилительных каскадов*, исключая возможность полезного использования выхода. Центральная область логистической функции, имеющая большой коэф-

фициент усиления, решает проблему обработки слабых сигналов, в то время как области с падающим усилением на положительном и отрицательном концах подходят для больших возбуждений. Таким образом, нейрон функционирует с большим усилением в широком диапазоне уровня входного сигнала.

Другой широко используемой активационной функцией является *гиперболический тангенс.* По форме она сходна с логистической функцией и часто используется биологами в качестве математической модели активации нервной клетки. В качестве активационной функции ИНН она записывается следующим образом:

OUT = th(x).

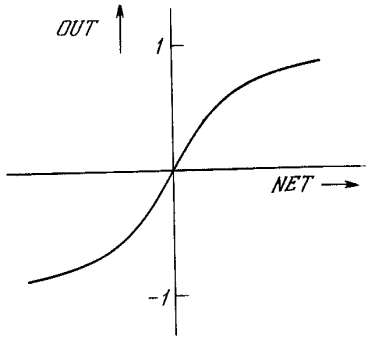


Рис. 1.5. Функция гиперболического тангенса

Подобно логистической функции гиперболический тангенс является S-образной функцией, но он симметричен относительно начала координат, и в точке NET = 0 значение выходного сигнала OUT равно нулю (см. рис. 1.5). В отличие от логистической функции гиперболический тангенс принимает значения различных знаков, что оказывается выгодным для ряда сетей.

Рассмотренная простая модель искусственного нейрона игнорирует многие свойства своего биологического двойника. Например, она не принимает во внимание задержки во времени, которые воздействуют на динамику системы. Входные сигналы сразу же порождают выходной сигнал. И, что более важно, она не учитывает воздействий функции частотной модуляции или синхронизирующей функции биологического нейрона, которые ряд исследователей считают решающими.

Несмотря на эти ограничения, сети, построенные из этих нейронов, обнаруживают свойства, сильно напоминающие биологическую систему. Только время и исследования смогут ответить на вопрос, являются ли подобные совпадения случайными или следствием того, что в модели верно схвачены важнейшие черты биологического нейрона.

3. Однослойные искусственные нейронные сети

Хотя один нейрон и способен выполнять простейшие процедуры распознавания, сила нейронных вычислений проистекает от соединений нейронов в сетях.

Простейшая сеть состоит из группы нейронов, образующих слой, как показано в правой части рис. 6.

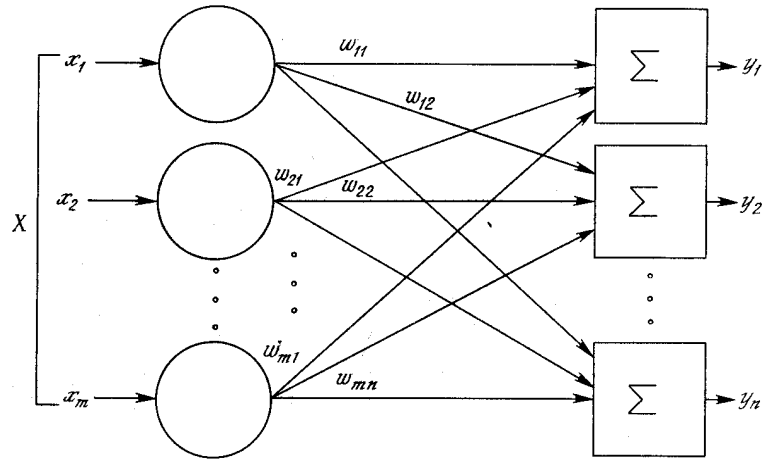


Рис. 6. Однослойная нейронная сеть

Отметим, что вершины-круги слева служат лишь для распределения входных сигналов. Они не выполняют каких-либо вычислений, и поэтому не будут считаться слоем. По этой причине они обозначены кругами, чтобы отличать их от вычисляющих нейронов, обозначенных квадратами. Каждый элемент из множества входов **X** отдельным весом соединен с каждым искусственным нейроном. А каждый нейрон выдает взвешенную сумму входов в сеть. В искусственных и биологических сетях многие соединения могут отсутствовать, все соединения показаны в целях общности. Могут иметь место также соединения между выходами и входами элементов в слое.

Удобно считать веса элементами матрицы **W**. Матрица имеет *m* строк и *n* столбцов, где m – число входов, а n – число нейронов. Например, $w_{2,3}$ – это вес, связывающий третий вход со вторым нейроном. Таким образом, вычисление *выходного вектора* **N**, компонентами которого являются выходы OUT нейронов, сводится к матричному умножению ***N = XW, где N и X – векторы-строки.***

4. Многослойные искусственные нейронные сети

Более крупные и сложные нейронные сети обладают, как правило, и большими вычислительными возможностями. Хотя созданы сети всех конфигураций, какие только можно себе представить, послойная организация нейронов копирует слоистые структуры определенных отделов мозга. Оказалось, что такие многослойные сети обладают большими возможностями, чем однослойные, и в последние годы были разработаны алгоритмы для их обучения.

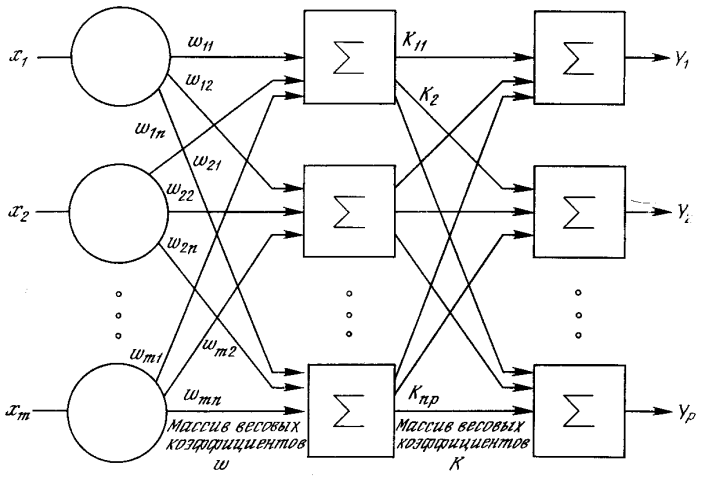


Рис. 7. Двухслойная нейронная сеть

Многослойные сети могут образовываться каскадами слоев. Выход одного слоя является входом для последующего слоя. Подобная сеть показана на рис. 7 и снова изображена со всеми соединениями.

Нелинейная активационная функция

Многослойные сети не могут привести к увеличению вычислительной мощности по сравнению с однослойной сетью. Это возможно лишь в том случае, если активационная функция между слоями будет нелинейной.

Вычисление выхода слоя заключается в умножении входного вектора на первую весовую матрицу с последующим умножением (если отсутствует нелинейная активационная функция) результирующего вектора на вторую весовую матрицу $(XW_1)W_2$.

Так как умножение матриц ассоциативно, то $X(W_1W_2)$.

Это показывает, что двухслойная линейная сеть эквивалентна одному слою с весовой матрицей, равной произведению двух весовых матриц. Следовательно, любая многослойная линейная сеть может быть заменена эквивалентной однослойной сетью. Известно, что однослойные сети весьма ограниченны по своим вычислительным возможностям. Таким образом, для расширения возможностей сетей по сравнению с однослойной сетью необходима нелинейная активационная функция.

Сети с обратными связями

У сетей, рассмотренных до сих пор, не было обратных связей, т. е. соединений, идущих от выходов некоторого слоя к входам этого же слоя или предшествующих слоев. Этот специальный класс сетей, называемых *сетями без обратных связей или сетями прямого распространения*, представляет интерес и широко используется.

Сети более общего вида, имеющие соединения от выходов к входам, называются *сетями с обратными связями*.

У сетей без обратных связей нет памяти, их выход полностью определяется текущими входами и значениями весов.

В некоторых конфигурациях сетей с обратными связями предыдущие значения выходов возвращаются на входы; выход, следовательно, определяется как текущим входом, так и предыдущими выходами. По этой причине сети с обратными связями могут обладать свойствами, сходными с кратковременной человеческой памятью, сетевые выходы частично зависят от предыдущих входов.

5. ТЕРМИНОЛОГИЯ, ОБОЗНАЧЕНИЯ И СХЕМАТИЧЕСКОЕ ИЗОБРАЖЕНИЕ ИНС

К сожалению, для ИНС еще нет опубликованных стандартов и устоявшихся терминов, обозначений и графических представлений. Порой идентичные сетевые парадигмы, представленные различными авторами, покажутся далекими друг от друга. Рассмотрим наиболее широко используемые термины.

Терминология

Многие авторы избегают термина «нейрон» для обозначения искусственного нейрона, считая его слишком грубой моделью своего биологического прототипа. Здесь термины «нейрон», «клетка», «элемент» используются взаимозаменяемо для обозначения «искусственного нейрона» как краткие и саморазъясняющие.

Дифференциальные уравнения или разностные уравнения

Алгоритмы обучения, как и вообще ИНС, могут быть представлены как в дифференциальной, так и в конечно-разностной форме. При использовании дифференциальных уравнений предполагают, что процессы *непрерывны* и осуществляются подобно большой аналоговой сети.

Для биологической системы, рассматриваемой на микроскопическом уровне, это не так. *Активационный уровень биологического нейрона определяется средней скоростью, с которой он посылает дискретные потенциальные импульсы по своему аксону.* Средняя скорость обычно рассматривается как аналоговая величина, но важно не забывать о действительном положении вещей.

Если моделировать искусственную нейронную сеть на аналоговом компьютере, то весьма желательно использовать представление с помощью дифференциальных уравнений. Однако сегодня большинство работ выполняется на цифровых компьютерах, что заставляет отдавать предпочтение *конечно-разностной форме* как наиболее легко программируемой. По этой причине обычно используются конечно-разностное представление.

Графическое представление

Как видно из публикаций, нет общепринятого способа подсчета числа слоев в сети. Многослойная сеть состоит, как показано на рис. 1.6, из чередующихся множеств нейронов и весов. Ранее в связи с рис. 1.5 уже говорилось, что входной слой не выполняет суммирования. Эти нейроны служат лишь в качестве разветвлений для первого множества весов и не влияют на вычислительные возможности сети. По этой причине первый слой не принимается во внимание при подсчете слоев, и сеть, подобная изображенной на рис. 1.6, считается двухслойной, так как только два слоя выполняют вычисления. Далее, веса слоя считаются связанными со следующими за ними нейронами. Следовательно, слой состоит из множества весов со следующими за ними нейронами, суммирующими взвешенные сигналы.

6. ОБУЧЕНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Среди всех интересных свойств ИНС ни одно не захватывает так воображения, как их способность к обучению. Их обучение до такой степени напоминает процесс интеллектуального развития человеческой личности, что может показаться, что достигнуто глубокое понимание этого процесса.

Но возможности обучения ИНС ограниченны, и нужно решить много сложных задач, чтобы определить, на правильном ли пути мы находимся. Тем не менее, уже получены убедительные достижения, такие как «говорящая сеть» Сейновского, и возникает много других практических применений.

Цель обучения

Сеть обучается, чтобы для некоторого множества входов давать желаемое (или, по крайней мере, сообразное с ним) множество выходов. Каждое такое входное (или выходное) множество рассматривается как вектор. Обучение осуществляется путем последовательного предъявления входных векторов с одновременной подстройкой весов в соответствии с определенной процедурой. В процессе обучения веса сети постепенно становятся такими, чтобы каждый входной вектор вырабатывал выходной вектор.

Обучение с учителем

Различают алгоритмы обучения с учителем и без учителя.

Обучение с учителем предполагает, что для каждого входного вектора существует целевой вектор, представляющий собой требуемый выход. Вместе они называются обучающей парой. Обычно сеть обучается на некотором числе таких обучающих пар. Предъявляется выходной вектор, вычисляется выход сети и сравнивается с соответствующим целевым вектором, разность (ошибка) с помощью обратной связи подается в сеть и веса изменяются в соответствии с алгоритмом, стремящимся минимизировать ошибку. Векторы обучающего множества предъявляются последовательно, вычисляются ошибки и веса подстраиваются для каждого вектора до тех пор, пока ошибка по всему обучающему массиву не достигнет приемлемо низкого уровня.

Обучение без учителя

Несмотря на многочисленные прикладные достижения, обучение с учителем критиковалось за свою биологическую неправдоподобность. Трудно вообразить обучающий механизм в мозге, который бы сравнивал желаемые и действительные значения выходов, выполняя коррекцию с помощью обратной связи. Если допустить подобный механизм в мозге, то откуда тогда возникают желаемые выходы?

Обучение без учителя является намного более правдоподобной моделью обучения в биологической системе. Развитая Кохоненом [3] и многими другими, она не нуждается в целевом векторе для выходов и, следовательно, не требует сравнения с предопределенными идеальными ответами.

Обучающее множество состоит лишь из входных векторов. Обучающий алгоритм подстраивает веса сети так, чтобы получались согласованные выходные векторы, т. е. *чтобы предъявление достаточно близких входных векторов давало одинаковые выходы.*

Процесс обучения, следовательно, выделяет статистические свойства обучающего множества и *группирует сходные векторы в классы*. Предъявление на вход вектора из данного класса даст определенный выходной вектор, но до обучения невозможно предсказать, какой выход будет производиться данным классом входных векторов. Следовательно, выходы подобной сети должны трансформироваться в некоторую понятную форму, обусловленную процессом обучения.

Это не является серьезной проблемой. Обычно не сложно идентифицировать связь между входом и выходом, установленную сетью.

Алгоритмы обучения

Большинство современных алгоритмов обучения выросло из *концепций Хэбба* [2]. Им предложена модель обучения без учителя, в которой *синаптическая сила (вес) возрастает, если активированы оба нейрона, источник и приемник.* Таким образом, часто используемые пути в сети усиливаются, чем и объясняется феномен привычки и обучения через повторение.

В ИНН, использующей *обучение по Хэббу*, наращивание весов определяется произведением уровней возбуждения передающего и принимающего нейронов. Это можно записать как

$$w_{ij}(n+1) = w(n) + \alpha OUT_i\, OUT_j,$$

где $w_{ij}(n)$ – значение веса от нейрона *i* к нейрону *j* до подстройки, $w_{ij}(n+1)$ – значение веса от нейрона *i* к нейрону *j* после подстройки, α – *коэффициент скорости обучения*, OUT_i – выход нейрона *i* и вход нейрона *j*, OUT_j – выход нейрона *j*.

Сети, использующие обучение по Хэббу, конструктивно развивались, однако за последние 20 лет были развиты более эффективные алгоритмы обучения. В

частности, в работах [4 – 6] и многих других были развиты алгоритмы *обучения с учителем*, приводящие к сетям с более широким диапазоном характеристик обучающих входных образов и большими скоростями обучения, чем использующие простое обучение по Хэббу.

В настоящее время используется огромное разнообразие обучающих алгоритмов. Потребовалась бы весьма значительное время для рассмотрения этого предмета полностью. Чтобы рассмотреть этот предмет систематически, если и не исчерпывающе, в каждой из последующих глав подробно описаны алгоритмы обучения для рассматриваемой в главе парадигмы.

Заключение

В последующих лекциях представлены и проанализированы некоторые наиболее важные сетевые конфигурации и их алгоритмы обучения.

Приведенные парадигмы дают представление об искусстве конструирования сетей в целом, его прошлом и настоящем. Многие другие парадигмы при тщательном рассмотрении оказываются лишь их модификациями. Сегодняшнее развитие нейронных сетей скорее эволюционно, чем революционно. Поэтому понимание представленных в данной книге парадигм позволит следить за прогрессом в этой быстро развивающейся области.

Упор сделан на интуитивные и алгоритмические, а не математические аспекты. Материалы адресованы скорее пользователю ИНС, чем теоретику. Сообщается, следовательно, достаточно информации, чтобы дать студенту возможность понимать основные идеи. Те, кто знаком с программированием, смогут реализовать любую из этих сетей. Сложные математические выкладки опущены, если только они не имеют прямого отношения к реализации сети. Приводятся ссылки на более строгие и полные работы.

Литература

1. Grossberg S. 1973. Contour enhancement, short-term memory, and consistencies in reverberating neural networks. Studies in Applied Mathematics 52:217,257.

2. Hebb D. 0. 1961. Organization of behavior. New York: Science Edition.

3. Kohonen T. 1984. Self-organization and associative memory. Series in Information Sciences, vol. 8. Berlin: Springer Verlag.

4. Rosenblatt F. 1962. Principles of neurodynamics. New York: Spartan Books. (Русский перевод: Розенблатт Ф. Принципы нейродинамики. – М.: Мир., 1965.)

5. Widrow B. 1959. Adaptive sampled-data systems, a statistical theory of adaptation. 1959 IRE WESCON Convention Record, part 4, pp. 88-91. New York: Institute of Radio Engineers.

6. Widrow B., Hoff M. 1960. Adaptive switching circuits. I960 IRE WESCON Convention Record, pp. 96-104. New York: Institute of Radio Engineers.

ПРИЛОЖЕНИЕ

ИНС СЕГОДНЯ

Имеется много впечатляющих демонстраций возможностей ИНС: сеть научили превращать текст в фонетическое представление, которое затем с помощью уже иных методов превращалось в речь; другая сеть может распознавать рукописные буквы; сконструирована система сжатия изображений, основанная на нейронной сети. Все они используют *сеть обратного распространения* – наиболее успешный, по-видимому, из современных алгоритмов. Обратное распространение, является систематическим методом для обучения многослойных сетей, и тем самым преодолевает ограничения, указанные Минским.

Как подчеркивается в следующих главах, обратное распространение не свободно от проблем. Прежде всего, *нет гарантии, что сеть может быть обучена за конечное время*. Много усилий, израсходованных на обучение, пропадает напрасно после затрат большого количества машинного времени. Когда это происходит, попытка обучения повторяется – без всякой уверенности, что результат окажется лучше. Нет также уверенности, что сеть обучится наилучшим возможным образом. Алгоритм обучения может попасть в «ловушку» так называемого локального минимума и будет получено худшее решение.

Разработано много других сетевых алгоритмов обучения, имеющих свои специфические преимущества. Некоторые из них обсуждаются в последующих главах. Следует подчеркнуть, что никакая из сегодняшних сетей не является панацеей, все они страдают от ограничений в своих возможностях обучаться и вспоминать.

ИНН уже продемонстрировали свою работоспособность, имеют уникальные потенциальные возможности, много ограничений и множество открытых вопросов. Такая ситуация настраивает на умеренный оптимизм. Авторы склонны публиковать свои успехи, но не неудачи, создавая тем самым впечатление, которое может оказаться нереалистичным.

Те, кто ищет капитал, чтобы рискнуть и основать новые фирмы, должны представить убедительный проект последующего осуществления и прибыли. Существует, следовательно, опасность, что ИНС начнут продавать раньше, чем придет их время, обещая функциональные возможности, которых пока невозможно достигнуть. Если это произойдет, то область в целом может пострадать от потери кредита доверия и вернется к застойному периоду семидесятых годов. Для улучшения существующих сетей требуется много основательной работы. Должны быть развиты новые технологии, улучшены существующие методы и расширены теоретические основы, прежде чем данная область сможет полностью реализовать свои потенциальные возможности.

ПЕРСПЕКТИВЫ НА БУДУЩЕЕ

ИНС предложены для задач, простирающихся от управления боем до присмотра за ребенком. Потенциальными приложениями являются те, где человеческий интеллект малоэффективен, а обычные вычисления трудоемки или неадекватны. Этот класс приложений во всяком случае не меньше класса, обслуживаемого обычными вычислениями, и можно предполагать, что ИНС займут свое место наряду с обычными вычислениями в качестве дополнения такого же объема и важности.

ИНС и экспертные системы

В последние годы над ИНН доминировали логические и символьно-операционные дисциплины. Например, широко пропагандировались *экспертные системы*, у которых имеется много заметных успехов, так же, как и неудач. Многое свидетельствует о том, что ИНН будут существовать, объединяясь в системах, где каждый подход используется для решения тех задач, с которыми он лучше справляется.

Эта точка зрения подкрепляется тем, как люди функционируют в нашем мире. *Распознавание образов отвечает за активность, требующую быстрой реакции*. Так как действия совершаются быстро и бессознательно, то этот способ функционирования важен для выживания во враждебном окружении. Вообразите только, что было бы, если бы наши предки вынуждены были обдумывать свою реакцию на прыгнувшего хищника?

Когда наша система распознавания образов не в состоянии дать адекватную интерпретацию, вопрос передается в высшие отделы мозга. Они могут запросить добавочную информацию и займут больше времени, но качество полученных в результате решений может быть выше.

Можно представить себе искусственную систему, подражающую такому разделению труда. Искусственная нейронная сеть реагировала бы в большинстве случаев подходящим образом на внешнюю среду. Так как такие сети способны указывать доверительный уровень каждого решения, то сеть «знает, что она не знает» и передает данный случай для разрешения экспертной системе. Решения, принимаемые на этом более высоком уровне, были бы конкретными и логичными, но они могут нуждаться в сборе дополнительных фактов для получения окончательного заключения. Комбинация двух систем была бы более мощной, чем каждая из систем в отдельности, следуя при этом высокоэффективной модели, даваемой биологической эволюцией.

Соображения надежности

Прежде чем ИНС можно будет использовать там, где поставлены на карту человеческая жизнь или ценное имущество, должны быть решены вопросы, относящиеся к их надежности.

Подобно людям, структуру мозга которых они копируют, ИНС сохраняют в определенной мере непредсказуемость. Единственный способ точно знать выход состоит в испытании всех возможных входных сигналов.

В большой сети такая полная проверка практически неосуществима и должны использоваться статистические методы для оценки функционирования. В некоторых случаях это недопустимо. Например, что является допустимым уровнем ошибок для сети, управляющей системой космической обороны? Большинство людей скажет, любая ошибка недопустима, так как ведет к огромному числу жертв и разрушений. Это отношение не меняется от того обстоятельства, что человек в подобной ситуации также может допускать ошибки.

Проблема возникает из-за допущения полной безошибочности компьютеров. Так как ИНС иногда будут совершать ошибки даже при правильном функционировании, то, как ощущается многими, это ведет к ненадежности – качеству, которое мы считаем недопустимым для наших машин.

Сходная трудность заключается в неспособности традиционных ИНС "объяснить", как они решают задачу. Внутреннее представление, получающееся в результате обучения, часто настолько сложно, что его невозможно проанализировать, за исключением самых простых случаев. Это напоминает нашу неспособность объяснить, как мы узнаем человека, несмотря на различие в расстоянии,

угле, освещении и на прошедшие годы. Экспертная система может проследить процесс своих рассуждений в обратном порядке, так что человек может проверить ее на разумность. Сообщалось о встраивании этой способности в ИНС, что может существенно повлиять на приемлемость этих систем.

ВЫВОДЫ

ИНС являются важным расширением понятия вычисления. Они обещают создание автоматов, выполняющих функции, бывшие ранее исключительной прерогативой человека. Машины могут выполнять скучные, монотонные и опасные задания, и с развитием технологии возникнут совершенно новые приложения.

Теория ИНС развивается стремительно, но в настоящее время она недостаточна, чтобы быть опорой для наиболее оптимистических проектов. В ретроспективе видно, что теория развивалась быстрее, чем предсказывали пессимисты, но медленнее, чем надеялись оптимисты, – типичная ситуация. Сегодняшний взрыв интереса привлек к нейронным сетям тысячи исследователей. Резонно ожидать быстрого роста нашего понимания ИНС, ведущего к более совершенным сетевым парадигмам и множеству прикладных возможностей.

ЛЕКЦИЯ 19

ИСКУССТВЕННАЯ НЕЙРОННАЯ СЕТЬ С ОБРАТНЫМ РАСПРОСТРАНЕНИЕМ ОШИБКИ

1. Краткие теоретические сведения

Пусть структура ИНС имеет вид, представленный на рис. 1.

Здесь $(x_1, x_2,\ldots,x_p)^T$ – входные сигналы, $w^1=(w_{11}^1, w_{22}^1,\ldots, w_{p1}^1)^T$, $w^2=(w_{11}^2, w_{22}^2,\ldots, w_{pk}^2)^T$ – весовые коэффициенты, соответственно, 1-го и 2-го уровней.

Каждый уровень ИНС имеет разное количество нейронов: A – p нейронов, S – l и R - k. Соотношение (p, l, k) первоначально неизвестно.

Начальные данные для обучение ИНС даны в табл. 1.

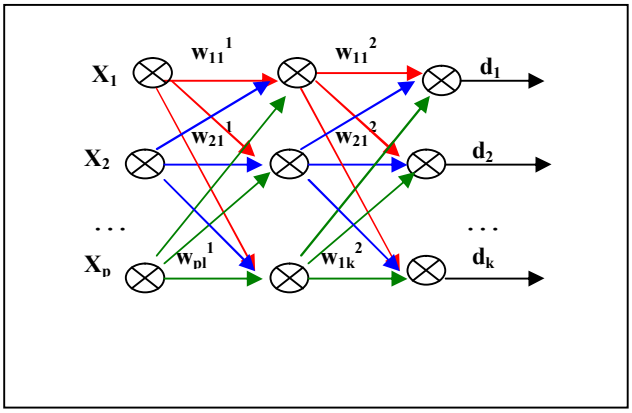


Рис. 1. Структура двухслойного перцептрона

Пусть N – кол-во точек входа и выхода, полученных в результате эксперимента или моделированием. $X=(X_1, X_2, \ldots, X_p)$– вектор входа, $D=(d_1, d_{2,\ldots,} d_k)$ - реальные или расчетные выходы.

Целевая функция, подлежащая минимизации:

$$E(\varpi)=\frac{1}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{l}(y_{ij}^1-d_{ij}^1)+\sum_{i=1}^{N}\sum_{j=1}^{k}(y_{ij}^2-d_{ij}^2)\right]^2=\frac{1}{2}\sum_{i=1}^{N}(\sum_{j=1}^{l}(y_{ij}^1-d_{ij}^1)+\sum_{j=1}^{k}(y_{ij}^2-d_{ij}^2))^2.$$

X_1	x_1^1	x_1^2	x_1^3	…	x_1^N
X_2	x_2^1	x_2^2	x_2^3	…	X_2^N
…	…	…	…	…	…
X_p	x_p^1	x_p^2	x_p^3	…	x_p^N
d_1	d_1^1	d_1^2	d_1^3	…	d_1^N
d_2	d_2^1	d_2^2	d_2^3	…	d_2^N
…	…	…	…	…	…
d_k	d_k^1	d_k^2	d_k^3	…	d_k^N

Поскольку значения выхода первого уровня d_{ij}^1 неизвестны, ограничимся выходом уровня R:

$$E(\varpi)=\frac{1}{2}\sum_{i=1}^{N}(\sum_{j=1}^{k}y_{ij}^2-d_{ij}^2)^2\to\min.$$

Используем метод градиентного спуска, что означает настройку весовых коэффициентов следующим образом:

$$\Delta\varpi_{ij}^{(n)}=-\eta\frac{\partial E}{\partial\omega_{ij}},\ n=1,2,$$

где $\varpi_{ij}^{(n)}$ - вес связи i-го нейрона (n-1)-го уровня с j-м нейроном n-го уровня, $1<0<\eta$ - коэффициент скорости обучения. Известно, что

$$\frac{\partial E}{\partial\omega_{ij}}=\frac{\partial E}{\partial y_j}\cdot\frac{\partial y_j}{\partial S_j}\cdot\frac{\partial S_j}{\partial\omega_{ij}},$$

где y_j – выход нейрона, S_j – взвешенная сумма его входных сигналов (аргумент активационной функции).

Третий множитель $y_i^{(n-1)}=\frac{\partial S_j}{\partial\omega_{ij}}$ - выход нейрона предыдущего уровня.

Первый множитель разложим следующим образом:

$$\frac{\partial E}{\partial y_i} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial S_k} \cdot \frac{\partial S_k}{\partial y_i} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{dy_k}{dS_k}.$$

Последняя сумма ищется среди нейронов (n-1)-го уровня. Введем новую замену:

$$\delta_j^{(n)} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial S_j},$$

получим рекурсивную формулу:

$$\delta_j^{(n)} = [\sum_k \delta_k^{(n+1)} \cdot \omega_{jk}^{(n+1)}] \cdot \frac{\partial y_j}{\partial S_j},$$

что дает возможность, зная $\delta_j^{(n+1)}$, вычислить $\delta_j^{(n)}$. Для выходного уровня

$$\delta_e^{(n)} = (y_e^{(n)} - d_e) \cdot \frac{dy_e}{dS_e},$$

Типичные функции активации сигмоид

$$y = \frac{1}{1 + e^{-\sum \omega_i x_i}}$$

или гиперболический тангенс

$$y = th\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad th'\, x = \frac{1}{ch^2(x)},$$

где ch(x) – гиперболический косинус,

Для случая сигмоида

$$\delta_e^{(n)} = (y_e^{(n)} - d_e) \cdot (1 - S_e) \cdot S_e. \quad (A)$$

для гиперболического тангенса

$$\delta_e^{(n)} = (y_e^{(n)} - d_e) \cdot (1 - S_e^2), \quad (B)$$

Тогда настройка весовых коэффициентов будет иметь вид

$$\Delta w_{ij}^{(n)} = -\eta \cdot \delta_j^{(n)} y_i^{(n-1)}, \; n = 1,2. \quad (C)$$

Для придания процессу коррекции весов некоторой инерциальности, чтобы сгладить резкие скачки при перемещении по поверхности целевой функции, последнее выражение дополняется значениями измененных весов на предыдущей итерации

$$\Delta w_{ij}^{(n)}(t) = -\eta \cdot (\mu \cdot \Delta \omega_j^{(n)} \cdot (t-1) + (1-\mu) \cdot \delta_j^{(n)} y_i^{(n-1)}), \; n = 1,2, \quad (D)$$

где μ -коэффициент инерционности, t – номер текущей итерации.

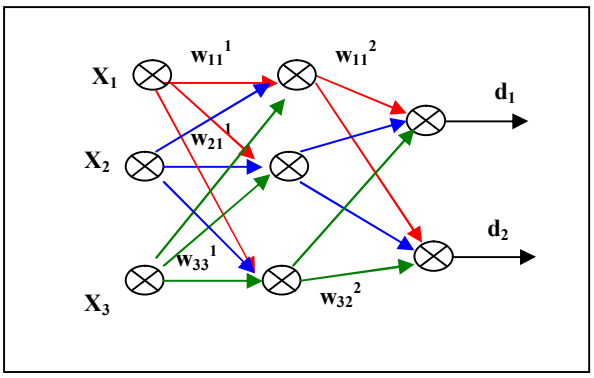


Рис. 2. Пример двухслойного перцептрона

2. Алгоритм обучения сети

Рассмотрим пример для p=3, l=3, k=2.

1. Подаем на вход сети первый образ (вар. 1 – (x_1=1, x_2=2, x_3=3). Полагаем начальные веса равными 0,5.

Рассчитываем взвешенную сумму сигналов, поступающих на нейроны 1-го уровня:

$$S_j^{(1)} = \sum_{i=o}^{3} \omega_{ij}^1 x_i^1, \quad j = 1,...,3$$

где 3 – число нейронов в уровне А, учитывая, что нулевой нейрон с постоянным выходным значением 1, что задает смещение $y_j^{(n-1)} = x_{ij}^{(n)}$ - I-ый вход j-го нейрона уровня n. Положим $y_1^0 = x_1^1, \; y_2^0 = x_2^1, \; y_3^0 = x_3^1$. В нашем случае

$$S_1^1 = w_{01}^1 + w_{11}^1 y_1^0 + w_{21}^1 y_2^0 + w_{31}^1 y_3^0,$$

$$S_2^1 = w_{02}^1 + w_{12}^1 y_1^0 + w_{22}^1 y_2^0 + w_{32}^1 y_3^0,$$

$S_3^1 = w_{03}^1 + w_{13}^1 y_1^0 + w_{23}^1 y_2^0 + w_{33}^1 y_3^0.$

2. Вычислим (например, для сигмоида) выход первого уровня:

$$y_1^1 = \frac{1}{1+e^{-S1}}, \quad y_2^1 = \frac{1}{1+e^{-S2}}, \quad y_3^1 = \frac{1}{1+e^{-S3}}.$$

3. Далее повторим процедуру для второго уравнения. При этом уравнения будут те же , но верхний индекс «1» нужно заменить на «2». Найдем $y_i^2, \quad i = 1,2.$

4. Истинные значения выхода сети $d_1^{(1)}$ и $d_2^{(1)}$ находятся из заданного вариантом выражения, которое сеть должна имитировать. Например, для первого варианта,

$d_1^{(1)} = x_1^2 - x_2^2 + x_3^2$;

$d_2^{(1)} = 1$, если $d_1^{(1)} - d_1^{(1)}$ расчет больше заданного порога β и $d_2^{(1)} = 0$ в противоположном случае.

5. Если ошибка сети превышает заданный порог β, т.е. то начинается процесс коррекции. Иначе – конец вычислений.

6. Вычислим корректирующие коэффициенты $\delta^{(n)}$ для сигмоидальной функции активации

$\delta_1^{(2)} = (y_1^{(2)} - d_1) \cdot (1 - S_1^2).$

$\delta_2^{(2)} = (y_2^{(2)} - d_2) \cdot (1 - S_2^2).$

7. Рассчитаем с помощью (C) или (D) измененные веса. В нашем случае, используя (C), имеем

$\Delta\varpi_{11}^2 = -\eta \cdot \delta_1^{(2)} y_1^{(1)}, \ \Delta\varpi_{12}^2 = -\eta \cdot \delta_2^{(2)} y_1^{(1)}, \ \Delta\varpi_{21}^2 = -\eta \cdot \delta_1^{(2)} y_2^{(1)}, \ \Delta\varpi_{22}^2 = -\eta \cdot \delta_2^{(2)} y_2^{(1)},$

$\Delta\varpi_{31}^2 = -\eta \cdot \delta_1^{(2)} y_{31}^{(1)}, \ \Delta\varpi_{32}^2 = -\eta \cdot \delta_2^{(2)} y_3^{(1)}.$

8. Скорректировать все веса нейронной сети

$w_{ij}^{(n)}(f) = w_{ij}^{(n)}(t-1) + \Delta w_{ij}^{(n)}(t).$

9. Входной образ изменяется на (+/-) 1 и переходим к новой итерации к п.1.

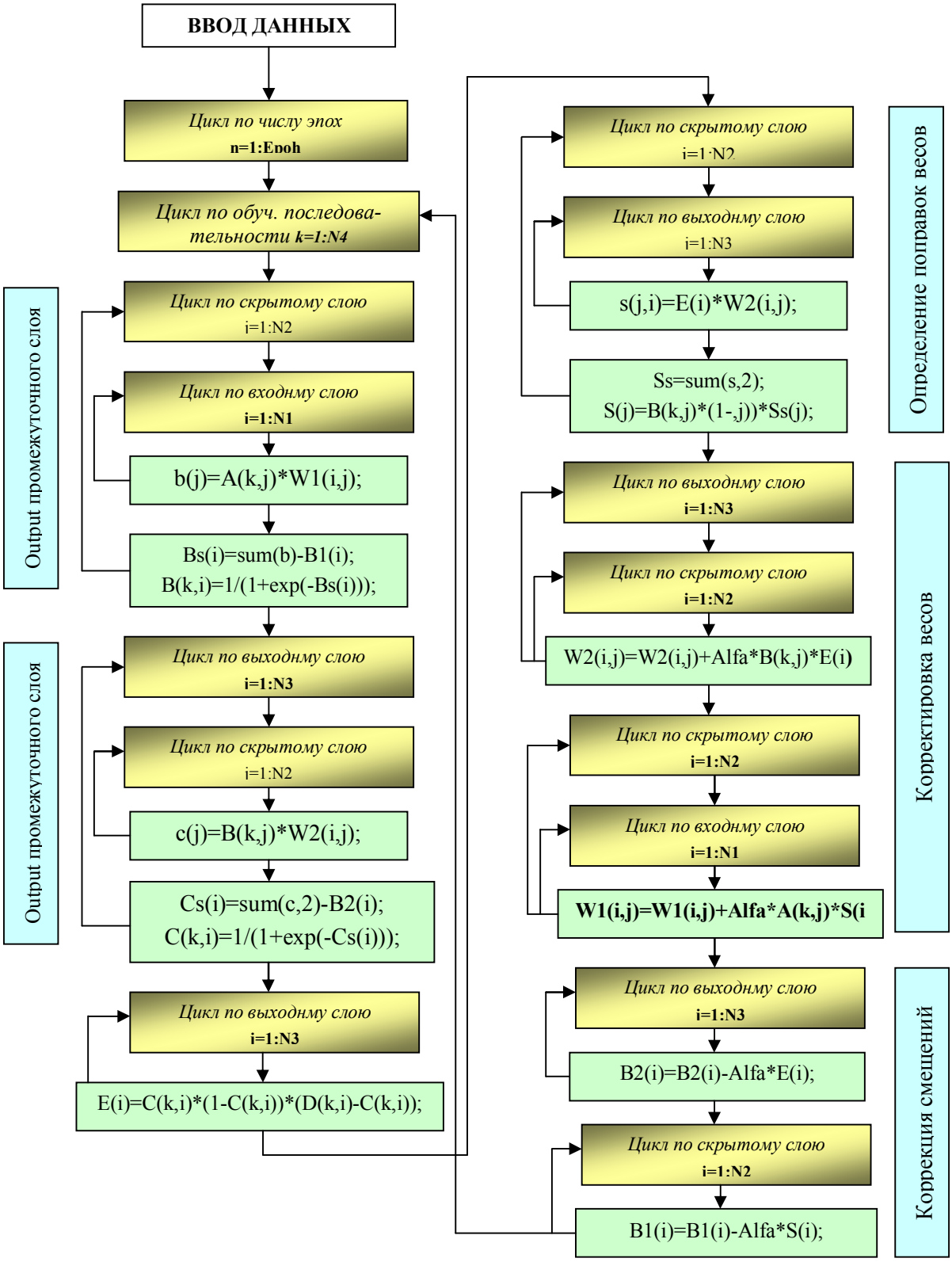
3. Структурная схема программы обучения ИНС с обратным распространением ошибки

Структурная схема программы обучения ИНС с обратным распространением ошибки представлена на рис. 3.

Контрольные вопросы

1. В чем состоит смысл алгоритма с обратным распространением ошибки?
2. Под каким оптимизационным методом функционирует алгоритм?
3. Какое значение имеет для алгоритма функция ошибки?
4. Для решения каких задач используется данный алгоритм?
5. Как перевести сочетание back propagation?
6. Назовите особенности процесса обучения НС с учителем,
7. Какие бывают методы обучения НС?
8. К какому классу обучения относится метод обратного распространения ошибки?
9. Почему порядок поступления приложений в начальной выборке может влиять на качество обучения?

10. Какими свойствами должна обладать функция активации при использовании в алгоритме обратного распространения ошибки?



ЛЕКЦИЯ 20

САМООРГАНИЗУЮЩИЕСЯ КАРТЫ КОХОНЕНА

1. Краткие теоретические сведения

Самоорганизующаяся карта Кохонена (*Self-organizing map* — SOM) — нейронная сеть с обучением без учителя, выполняющая задачу визуализации и кластеризации. Идея сети предложена финским учёным Т. Кохоненом. Является методом проецирования многомерного пространства в пространство с более низкой размерностью (чаще всего, двумерное), применяется также для решения задач моделирования, прогнозирования и др.

SOM Кохонена используются для решения таких задач, как моделирование, прогнозирование, выявление наборов независимых признаков, сжатие информации, а также для поиска закономерностей в больших массивах данных. Наиболее часто описываемый алгоритм применяется для кластеризации данных.

SOM состоит из компонентов, называемых *узлами* или *нейронами*. Их количество задаётся аналитиком, исходя из особенностей решаемой задачи.

Каждый из узлов описывается двумя векторами. Первый — т. н. *вектор веса m*, имеющий такую же размерность, что и входные данные. Второй — вектор *r*, представляющий собой *координаты узла на карте*.

Карта Кохонена визуально отображается с помощью ячеек прямоугольной или шестиугольной формы; последняя применяется чаще, поскольку в этом случае расстояния между центрами смежных ячеек одинаковы, что повышает корректность визуализации карты.

Изначально известна размерность входных данных, по ней некоторым образом строится первоначальный вариант карты. В процессе обучения векторы веса узлов приближаются к входным данным. Для каждого наблюдения (семпла) выбирается наиболее похожий по вектору веса узел, и значение его вектора веса приближается к наблюдению. Также к наблюдению приближаются векторы веса нескольких узлов, расположенных рядом. Таким образом, если в множестве входных данных два наблюдения были схожи, на карте им будут соответствовать близкие узлы. Циклический процесс обучения, перебирающий входные данные, заканчивается при достижении картой допустимой (заранее заданной) погрешности, или по совершении заданного количества итераций.

Таким образом, в результате обучения карта Кохонена классифицирует входные данные на кластеры и визуально отображает многомерные входные данные в двумерной плоскости, распределяя векторы близких признаков в соседние ячейки и раскрашивая их в зависимости от анализируемых параметров нейронов.

В результате работы алгоритма получаются следующие карты:

карта входов нейронов — визуализирует внутреннюю структуру входных данных путём подстройки весов нейронов карты. Обычно используется несколько карт входов, каждая из которых отображает один из них и раскрашивается в зависимости от веса нейрона. На одной из карт определенным цветом обозначают область, в которую включаются приблизительно одинаковые входы для анализируемых примеров;

карта выходов нейронов — визуализирует модель взаимного расположения входных примеров. Очерченные области на карте представляют собой кластеры, состоящие из нейронов со схожими значениями выходов;

специальные карты — это карта кластеров, полученных в результате применения алгоритма самоорганизующейся карты Кохонена, а также другие карты, которые их характеризуют.

Таким образом, алгоритм Кохонена дает возможность строить *искусственную нейронную сеть* (ИНС) позволяющую разделять вектора входных сигналов на подгруппы.

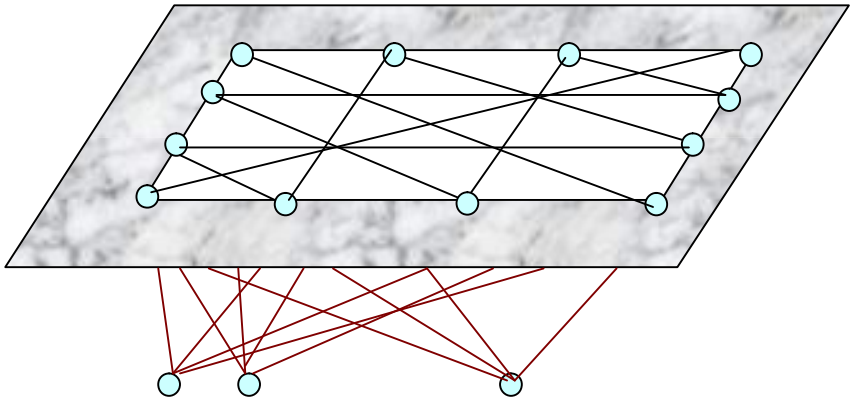


Рис 1. Карта Кохонена

ИНС формируется из *M* нейронов, образующих на плоскости прямоугольную решетку (рис.1). Входные сигналы подаются на всю ИНС. В процессе функционирования настраиваются синаптические веса нейронов.

Входные сигналы, образованные векторами вещественных чисел, последовательно поступают на вход сети. Желательные входы не назначаются. После того, когда на вход поступило достаточно количество векторов, синаптические веса определяют кластеры.

Кроме того, веса организуются так, что типологически близкие узлы чувствительны для схожих входных сигналов.

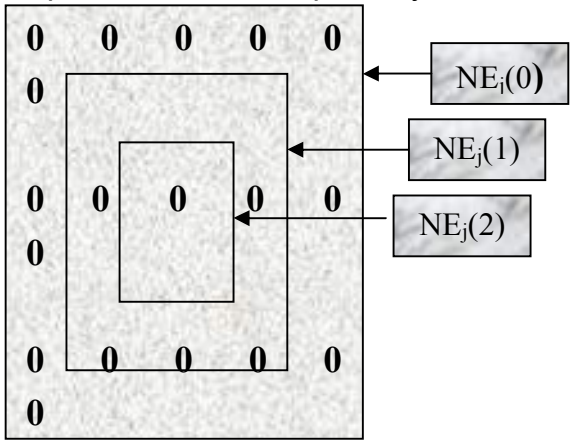


Рис.2. Зоны топологической близости

Для организации алгоритма необходимо назначить меру близости нейронов. На рис. 2 показаны зоны топологической близости нейронов на карте в разные моменты времени. $NE_j(t)$ – множество нейронов, которые предполагаются близкими к нейрону ***j*** в момент времени ***t***. Зоны близости со временем уменьшаются.

Весовым коэффициентам сети придают малые начальные значения. Начальное число весов равно $M \times N$.

2. АЛГОРИТМ КОХОНЕНА

Инициализация

Наиболее распространены три способа задания первоначальных весов узлов:

- Задание всех координат случайными числами.

- Присваивание вектору веса значение случайного наблюдения из входных данных.

Выбор векторов веса из линейного пространства, натянутого на главные компоненты набора входных данных

1. НС предъявляется новый входной сигнал;
2. Вычисляются евклидово расстояния до всех нейронов

$$d_j = \sum_{i=0}^{N-1} [x_i(t) - w_{ij}(t)]^2, \quad (1)$$

где $x_i(t)$ – i-й элемент входного сигнала в момент времени *t*, $w_{ij}(t)$ – вес связи *i*-го элемента входного сигнала с *j*-ым нейроном в момент времени t.

3. Выбирается нейрон с наименьшим расстоянием d_j.
4. Настраиваются весовые коэффициенты для *j*-го нейрона и всех нейронов из его зоны близости. Новые значения весов

$$w_{ij}(t+1) = w_{ij}(t) + v(t)[x_i(t) - w_{ij}(t)], \quad (2)$$

где *v(t)* – шаг обучения, изменяющийся со временем (дополнительное число меньше единицы).

1. Переход на п. 2.

3. Особенности модели

Устойчивость к зашумленным данным, быстрое и неуправляемое обучение, возможность упрощения многомерных входных данных с помощью визуализации.

Самоорганизующиеся карты Кохонена могут быть использованы для кластерного анализа только в том случае, если заранее известно число кластеров.

Важным недостатком является то, что окончательный результат работы нейронных сетей зависит от начальных установок сети. С другой стороны, нейронные сети теоретически могут аппроксимировать любую непрерывную функцию, что позволяет исследователю не принимать заранее какие-либо гипотезы относительно модели.

Контрольные вопросы

1. Для решения каких задач используется карта Кохонена?
2. Сформулируйте содержание алгоритма Кохонена.
3. Назовите особенности модели карт Кохонена
4. Недостатки и достоинства карты Кохонена.

ЛИТЕРАТУРА:

1. T. Kohonen, Self-Organizing Maps (Third Extended Edition), New York, 2001, 501 pages.

2. Дебок Г., Кохонен Т. Анализ финансовых данных с помощью самоорганизующихся карт, Альпина Паблишер, 2001, 317 стр.

3. Зиновьев А. Ю. Визуализация многомерных данных. — Красноярск: Изд. Красноярского государственного технического университета, 2000. — 180 с.

4. Каллан Р. Основы концепции нейронных сетей / Пер. с англ. – М.: Изд. дом «Вильямс», 2001. – 288с.

ЛЕКЦИЯ 21

ЭВОЛЮЦИОННОЕ МОДЕЛИРОВАНИЕ

1. Эволюционное моделирование: Постановка задачи.

В основе технологии эволюционного моделирования, предложенной в [1] и изученной в большом числе современных работ [2-12], лежит идея замены прямой параметрической и структурной оптимизации модели динамической системы процессом последовательной эволюции изменяющихся моделей. Эволюция математических моделей осуществляется в соответствии с принципами дарвиновской теории развития - изменчивостью, селекцией и отбором.

Имитация эволюционной изменчивости осуществляется на основе случайных вариаций параметров модели. Обычно используются три основных вида таких модификаций.

1. Естественная изменчивость. Данный вид изменчивости представляет собой относительно небольшие случайные колебания параметров модели при переходе от одного поколения к другому.

2. Параметрическая мутация. Отличается от естественной изменчивости низкой вероятностью появления и крайне большим значением изменяемого параметра.

3. Непараметрическая мутация. Характеризуется изменением структуры модели. В простейшем случае может состоять в изменении порядка или размерности модели.

В результате процесса изменчивости происходит переход от группы *моделей-родителей* (МР) к следующему поколению *моделей-потомков* (МП). При этом количество потомков должно значимо превосходить количество МР из предыдущей генерации.

Вторым этапом (на каждой итерации) эволюционного моделирования являются процессы селекции и отбора. Новое поколение моделей проверяется по экзогенному критерию выживаемости, определяемому качеством решения функциональных задач. В соответствии с этим критерием осуществляется отбор наиболее эффективных моделей, которые переходят в категорию нового поколения МР и допускаются к очередным изменениям.

Заметим, что в процессе селекции и отбора «выживает» и формирует новое поколение не только наилучшая (среди предыдущего поколения) модель, а несколько моделей, в наибольшей степени удовлетворяющие критерию селекции. Такой подход, названный *принципом незаконченных решений*, позволяет реализовать постулат системного анализа, в соответствии с которым наилучшее решение формируется из последовательности шагов или элементов, не являющихся оптимальными на каждом (промежуточном) шаге итерации.

Данное отличие от традиционной схемы пошаговой оптимизации является существенным и является одним из характеристических свойств эволюционной оптимизации.

Следует заметить, что эволюционное моделирование, по сути, представляет собой процесс стохастической самоорганизации прикладной математической модели. При этом извне задается лишь критерии селекции. Процесс модификации моделей остается не контролируемым и формируется самой программой в соответствии с принятой технологией генерации поколений. По мнению авторов данного метода [1], такой подход позволяет по-новому подойти к задаче создания ис-

кусственного интеллекта, когда роль внешнего дополнения в теореме Геделя о неполноте будет выполнять программа, обладающая определенной свободой выбора.

2. Моделирование изменчивости. Выбор параметра, подлежащего модификации при переходе от МР к модели-потомку осуществляется случайным образом, путем розыгрыша номеров параметров МР, подлежащих модификации. Описание технологии розыгрыша существует во всех книгах, посвященных методу Монте-Карло, начиная с [13, 14].

Естественная изменчивость обычно формируется в виде розыгрыше случайной величины, подчиненной нормальному закону $\tilde{a}_{ik} \in N\{a_{ik}, \sigma_i^2\}$, где a_{ik} - значение i-го параметра в предшествующей k-й модели, σ_i - среднеквадратическое отклонение (*ско*) вариаций a_i.

Для нестационарной динамики *ско* постоянно изменяется. В этом случае новое значение изменяемого параметра можно формировать на основе эмпирического соотношения $a_{i(k+1)} = a_{ik} + u[a_{ik} + \varepsilon \cdot a_{ik},\ a_{ik} - \varepsilon \cdot a_{ik}]$, где $u[\cdot]$ - генератор равномерной случайной величины из диапазона, указанного в квадратных скобках, ε - величина, определяющая степень вариабельности модифицируемого параметра a_i. Обычно данная величина лежит в пределах 5-10% от собственного значения изменяемого параметра. Очевидно, что в последнем случае нужно контролировать принадлежности нового значения параметра диапазону допустимых изменений.

Параметрическая мутация отличается от естественной изменчивости низкой вероятностью появления (в пределах 1-5%) и крайне большим значением изменяемого параметра ε, лежащим за пределами 3σ для стационарных процессов или имеющих ε, превышающее 50% от значения изменяемого параметра для нестационарных процессов.

Для моделирования непараметрической мутации, необходимо *a'priori* иметь банк допустимых моделей, выбор из которого реализуется на основе случайного розыгрыша. Непараметрическая мутация реализуется с еще меньшей вероятностью, чем параметрическая, но значительно чаще, чем в природе. Обычно вероятность такой мутации лежит в пределах 0.1-1%.

Заметим, что решение задачи изменчивости можно решать различными методами. В природе в большинстве случаев формирование генома потомка осуществляется путем комбинирования генов из геномов разнополых родителей . Выбор параметров потомка осуществляется путем случайного выбора параметров от одного или другого родителя. Такой подход используется при использовании генетических алгоритмов последовательной оптимизации.

3. Особенности эволюционной оптимизации для хаотических процессов

Многочисленные приложения эволюционного моделирования и связанных с ним генетических алгоритмов [15-18] позволяют находить эффективные решения для нестационарных процессов, обладающих инерционностью или другими регуляризующими факторами. В отношении хаотических процессов трудно надеяться на то, что решение, эффективное на обучающей выборке, будет хотя бы в какой-то степени осмысленным на новых данных.

В отличие от эволюционного моделирования, эволюционную оптимизацию управляющих стратегий не интересует степень подобия используемой математической модели. Задача состоит в выборе такой модели, которая бы обеспечивала наибольшую эффективность реализации управляющей стратегии. Более того, сама стратегия, как совокупность решающих правил, может войти в состав изменяемых и модифицируемых категорий.

Простейшее решение состоит в формировании банка стратегий, из которого, в процессе непараметрических мутаций случайным образом выбираются наборы решающих правил, списки модифицируемых параметров, их критические значения и диапазоны изменений. Достоинством такого подхода является его реализуемость. Однако при этом ограничивается произвол машинного выбора, отсутствует возможность получения радикально новых стратегий, не предусмотренных программистом. Очевидно, что полное снятие ограничений при случайном формировании управляющих стратегий приведет к бесконечному количеству бессмысленных решающих правил. Ожидание появления сколько-либо разумного решения потребует времени, соизмеримого с реальной биологической эволюцией. В то же время любая регуляризация, любая совокупность ограничений может закрыть доступ к неожиданным оригинальным решениям. При этом остается открытым вопрос о технологии искусственной генерации вариантов управляющих стратегий.

Эволюционная технология, как и вся вероятностно-статистическая парадигма, ориентирована на комфортную гипотезу о повторяемости опытов в неизменных или медленно меняющихся условиях. Переход к нестационарным, а тем более, хаотическим процессам неизбежно разрушает все статистические технологии, в том числе и эволюционное моделирование. Однако в природе хаоса, как правило, присутствуют некоторые регуляризующие эффекты, снижающие степень тотальной неопределенности. Если эволюционная технология сможет выделить, хотя бы не явно, и использовать такие скрытые закономерности, то задача построения выигрышной стратегии может оказаться реализуемой.

Кроме того, применение эволюционной вычислительной схемы позволят ответить на вопрос о принципиальной допустимости той или иного класса управляющих стратегий.

4. Формализованная постановка эволюционной оптимизации.

Имеется некоторая игровая стратегия $S\{K, p^*, a\}$, определяемая критериальными правилами K, критическими значениями правил принятия решений p^* и технологическими параметрами алгоритма анализа данных a. Эффективность стратегии $Eff(S)$ оценивается на основе ее применения к временным рядам наблюдений $Y(t)$, образующим в совокупности опытный полигон ретроспективных данных.

Введем два нелинейных оператора.

1. Оператор изменчивости и размножения стратегий

$$\Phi(S): \quad S \Rightarrow \{S_1, ..., S_{N_a}: \quad S_i \neq S_j \neq S, \quad \forall i, j\}.$$

2. Оператор селекции и отбора

$$\Psi(S_1,...,S_{N_g}): \quad \{S_1, .., S_{N_g}\} \Rightarrow$$
$$\Rightarrow \{S_{<1)>}, ..., S_{<N_a>}: \quad Eff(S_{<1>}) \geq ...$$
$$... \geq Eff(S_{<N_a>}) \geq Eff(S_j), \quad \forall j > N_a\},$$

где N_a - количество «выживших» стратегий, которые допускаются для дальнейшего размножения-модификации (индекс *a* – от «ancestor», «предок»); $N_g = N_a(1 + N_d)$ - количество стратегий одного поколения, подлежащие селекции-отбору (индекс *g* – от «generation», «поколение»), N_d - количество стратегий-потомков, генерируемых в соответствии с правилами размножения-модификации на каждой итерации (индекс *d* – от «descendant», «потомок»).

Пусть $S_o = S\{K_0, p_0^*, a_0\}$ - конкретный вариант управляющей стратегии с заданными параметрами, принятой в качестве базовой «стратегии-родителя». Тогда технология эволюционной оптимизации сводится к циклическому повторению выполнения последовательности операторов

$$S_o \Rightarrow \Phi(S_0) = \{S_1,...,S_{N_g}\}$$
$$\Uparrow \qquad\qquad \Downarrow$$
$$\Psi(S_1,...,S_{N_g}) = S_0 = \{S_{<1>},...,S_{<N_a>}\}$$

Поскольку селекция осуществляется по критерию превосходства, оптимальность терминального решения не гарантируется. Однако оно будет наилучшим из всего множества случайного перебора, формируемого в процессе реализации эволюционной технологии.

5. Вычислительные аспекты. Рекуррентные операции завершаются по выполнению заданного числа итераций или при превышении показателя эффективности заданного порогового значения

Процесс эволюционной оптимизации является заведомо сходящимся к более эффективным стратегиям в силу самого его построения. Действительно, новое поколение всегда включает в свой состав и стратегии-родители, отобранные по критерию наибольшей эффективности. Таким образом, наиболее эффективные стратегии в принципе не могут быть отброшены принятой процедурой селекции и отбора. Однако высокую скорость сходимости ожидать не приходиться в силу случайности процесса модификации. Наиболее вероятно, что скорость сходимости будет близка к скорости сходимости случайного поиска и зависит от размера формируемого поколения N_g. Можно предположить, что скорость сходимости будет выше, если количество *стратегий-потомков* (СП) N_d сделать зависимым от эффективности *стратегий-родителей* (СР), т.е. $N_d = k(Eff(S_a)), \quad k \geq 1$. Иными словами, более эффективный предок может порождать большее количество потомства. Однако данное утверждение требует дополнительной проверки. Возможны и другие методы регуляризации, направленные на увеличение скорости сходимости эволюционной оптимизации.

Функциональная структура алгоритма эволюционной оптимизации управляющих стратегий приведена на рис. 1. Последовательность эволюции представлена схемой процесса, развивающегося снизу вверх. Вторые индексы у стратегий упущены, чтобы не загромождать и без того насыщенную схему.

Рассмотрим работу приведенного алгоритма. На первом шагу формируется некоторая базовая стратегия (прототип), структура которой выбирается случайно или исходя из имеющегося априорного опыта по управлению активами.

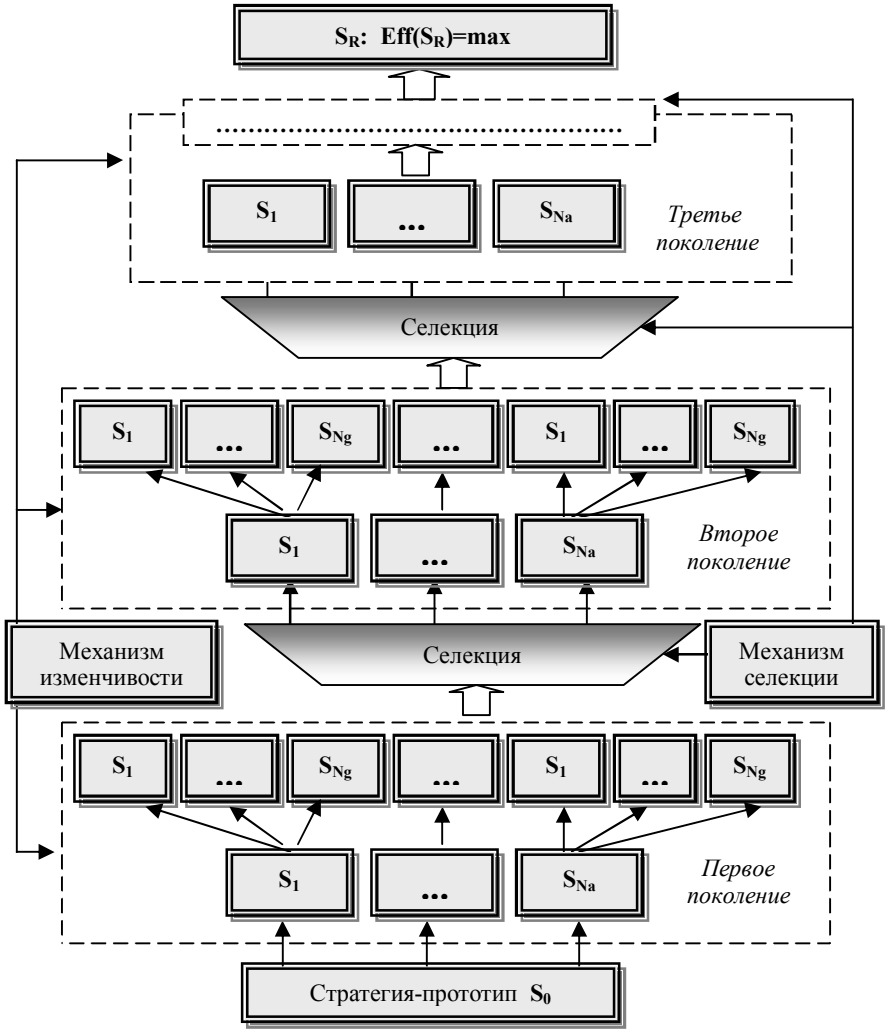


Рис. 1. Общая функциональная структура алгоритма эволюционной оптимизации управляющей стратегии

С помощью генератора изменчивости базовая стратегия-прототип видоизменяется, порождая N_a новых СР. Для этого разыгрывается тип изменения (естественная изменчивость, параметрическая или непараметрическая мутация) и в зависимости от сделанного выбора осуществляется разыгрывание объекта модификации и собственно величина соответствующего изменения.

В свою очередь каждая из новых стратегий является прототипом, случайные изменения которых порождают еще N_d новых стратегий-потомков.

Вся совокупность исходных стратегий образует первое поколение стратегий, подлежащих сравнительному анализу по критерию эффективности. Каждая из стратегий первого поколения проходит процедуру тестирования путем применения на множестве ретроспективных наблюдений $\{Y(t), Y(t-T)\}$, где T - размер испытательного полигона данных. Сравнение эффективности сформированных стратегий осуществляется путем прямой ранжировка ряда $Eff(S_i),\ \ i=0,...,N_g$, позволяющей отобрать заданное количество N_a «выживших» стратегий, которые допускаются для дальнейшего «размножения» (модификации). Индекс *a* – от «ancestor», «предок».

Отобранные стратегии являются родителями нового множества модифицированных стратегий-потомков и вместе с ними образуют второе поколение.

На втором шагу формируются N_d стратегий-потомков путем внесения различного рода изменений в N_a стратегий-предков. В результате образуется $N_g = N_a + N_d$ управляющих стратегий, эффективность каждой из которых $Eff(S_i), \quad i = 0,..., N_1$ вновь оценивается на том же временном ряду наблюдений $\{Y(t), Y(t-T)\}$. Каждая из отобранных СР порождает путем модификаций N_d стратегий-потомков (СП). Тогда общее количество стратегий нового поколения (*g*, generation) составит $N_g = N_a(1 + N_d).$ Последующие итерации повторяют последовательную работу механизмов изменчивости, селекции и отбора.

Процессы "Изменения-размножения" и "Селекции-отбора" стратегий повторяются в течении заданного числа поколений. Остановка цикла генерации поколений может быть проведена и раньше, например, на основе критерия превышения порога точности или критерия сходимости результатов прогноза на тестовой совокупности данных. В конечном счете, итерационная процедура позволяет выявить наилучшую стратегию, наиболее успешно функционирующую на заданном интервале наблюдений.

6. Пример. Идентификация поверхности отклика второго порядка. Имеется некоторая априорная нелинейная гиперповерхность отклика *M*-го порядка

$$y = y(x_1, x_2, ..., x_M) \tag{1}$$

и совокупность из *N* векторных наблюдений за каждым из *M* независимых параметров $Z_{<N:M>}$.

Осуществляется МНК аппроксимация этой гиперповерхности полиномиальной моделью не выше 2-го порядка. При этом

$$\hat{y} = f(x_1, ..., x_M,\ x_1^2, ..., x_M^2,\ x_1x_2, ..., x_1x_M, ..., x_{M-1}x_M). \tag{2}$$

Требуется построить наилучшую аппроксимацию, для которой

$$Q_N = \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 = \min. \tag{3}$$

Состав модели (2) может быть любым. Таким образом, речь идет о выборе наилучшей структуры аппроксимирующей модели.

Очевидно, что такой выбор можно сделать полным перебором переменных, входящих в (1). В правую часть (1) входит переменных N_{ALL}= *M* (число независимых переменных) + *M* (число квадратов независимых переменных) +*(M*M-M)/2* (число различных парных произведений независимых переменных).

Тогда полное число возможных моделей составит

$$N_{Мод} = C^1_{N_{ALL}} \cdot C^2_{N_{ALL}} \cdot ... \cdot C^{N_{ALL}-1}_{N_{ALL}} \cdot C^{N_{ALL}}_{N_{ALL}}.$$

Альтернативу к полному перебору дант технология эволюционного моделирования, имитирующая процесс биологической эволюционной оптимизации, основанной на принципах рандомизации, наследственности, селекции, отбора.

В соответствии с технологией эволюционного моделирования, поиск оптимальной модели (2) представлен в виде блок-схемы на рис. 1 и состоит в следующем:

1. Формируется полная модель наблюдений второго порядка с двумя переменными.

$$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2 + a_6x_1x_2.$$

Выбирается базовая модель (из таблицы 1 вариантов моделей, подлежащих идентификации).

2. Формируется первое поколение моделей-родителей, состоящее из $M_p = 3$ моделей путем случайного выбора значений $a_i,\ \ i = 1,...6$ из интервалов их допустимых вариаций $Da_i,\ \ i = 1,...6$.

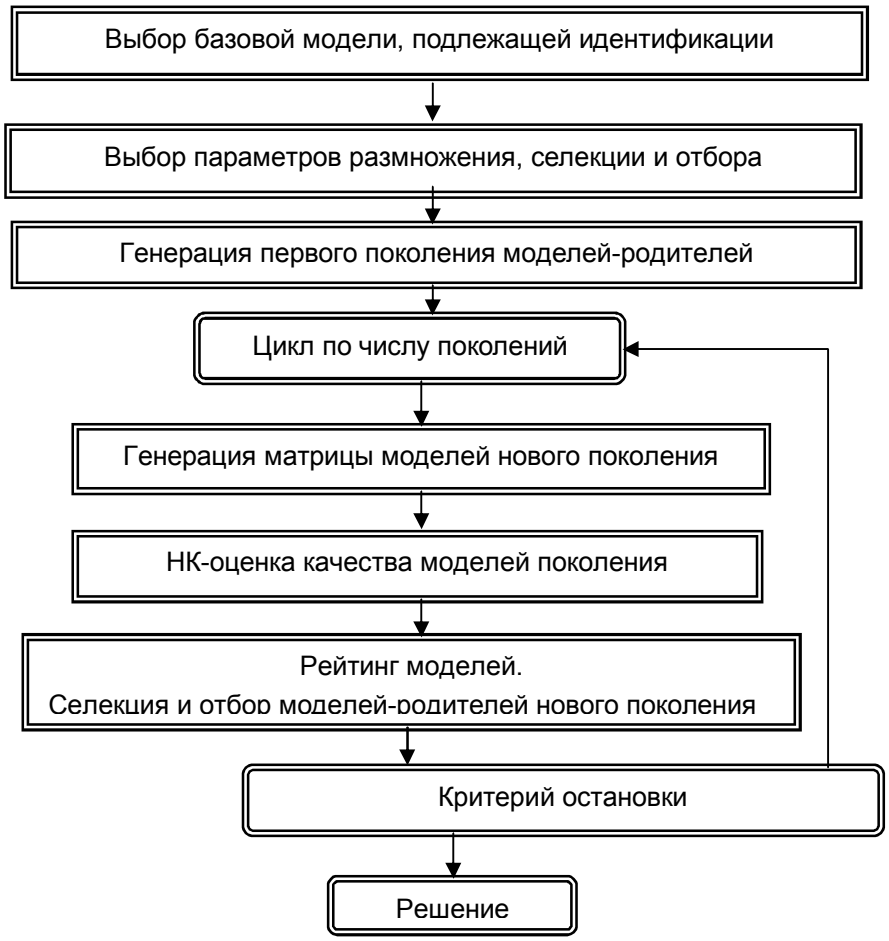


Рис. 1. Блок-схема решения задачи эволюционной идентификации

3. Формируется поколение моделей-потомков , путем внесения в каждую из моделей-родителей только по одному изменению. Каждай из моделей-родителей порождает 3 модели-потомка. Таким образом, формируется матрица моделей-потомков, состоящая из $M_d = M_p \cdot k_b = 3 \cdot 3 = 9$ моделей. Здесь $k_b = 3$ - коэффициент размножения моделей (kBreading).

Для формирования каждой модели потомка осуществляется следующая последовательность операций:

- выбор номера параметра модели, подлежайшего изменению, осуществляется случайным розыгрышем;
- разыгрывается один из трех возможных типов модификации:

с вероятностью $P_1 = 0.7$ осуществляется небольшая (в пределах 10% от исходного значения) параметрическая модификация одного из шести параметров модели;

с вероятностью $P_2 = 0.2$ осуществляется большая (в пределах 30% от исходного значения) параметрическая модификация одного из шести параметров модели (параметричсекая мутация);

с вероятностью $P_2 = 0.1$ осуществляется структурная мутация, т.е. данный параметр (ген) исключается из модели (или, наоборот, в слуцчае его отсутствия, добавляется в модель).

4. Матрица поколения *Generation* формируется путем объединения (вертикальная конкатенация) матрицы-родителей и матрицы-потомков.

5. Организуется процесс селекции и отбора. Простейший вариант селекции связан с использованием квадратичной метрики рассогласования между параметрами базовой модели и параметрами каждой модели (генома) из матрицы-поколения:

$$\Psi(i) = \sum_{k=1}^{m} (a_k - a_{ki})^2 .$$

Формируется рейтинг моделей поколения по критерию $\min\{\Psi(i)\}$. Первые *mSurvive* моделей выживают и образуют матрицу-родителей следующего поколения.

5. Осуществляется переход к п.3., формируется матрица следующего поколения и т.д.

Программа завершается по достижению заданного уровня Ψ_0 точности подгонки или по достижению заданной величины числа итераций (поколений) N_g.

Вопросы для самопроверки:

1. Для решения каких задач используется ЭМ?
2. Чем отличается ЭМ от случайного поиска.
3. Назовите критерии селекции.
4. Перечислите особенности реализации алгоритма ЭМ.
5. Каким образом реализуется процесс изменчивости наследования?
6. Каким образом реализуется процесс параметрической мутации наследования?

ЛИТЕРАТУРА

1. Fogel L.J., Owens A.J., Walsh M.J. Artificial intelligence through simulated evolution. / N.Y.: John Wiley & Sons. – 1966. – 231p.
2. Аверченков В.И. Эволюционное моделирование и его применение: монография / В.И. Аверченков, П.В. Казаков. 2-е изд., стереотип. — М.: ФЛИНТА. — 2011. — 200с.
3. Каширина И.Л. Эволюционное моделирование: учебное пособие для втузов. / Воронеж: Изд. центр ВГУ. — 2011. — 60с.
4. Курейчик В. Эволюционное моделирование и генетические алгоритмы. / В. Курейчик, Л. Гладков, В. Курейчик. — Lambert Academic Publishing. — 2011. — 260с.
5. Карпов В.Э. Методологические проблемы эволюционных вычислений // Искусственный интеллект и принятие решений. — 2012. — №4. — С.95-102.
6. Рутковский Л. Методы и технологии искусственного интеллекта. / М.: Горячая линия–Телеком. — 2010. — 520с.
7. Mukhopadhyay A. A. Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I / Mukhopadhyay A., Maulik U., Bandyopadhyay S., Coello C.A. IEEE Transactions on Evolutionary Computation. — 2014. — V.18. — N1. — P. 4-19.
8. Mukhopadhyay A. A. Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part II // Mukhopadhyay A., Maulik U., Bandyopadhyay S., Coello C.A. IEEE Transactions on Evolutionary Computation. — 2014. — V.18. — N1. — P. 20–35.

9. Carreno J. E. Multi-objective optimization by using evolutionary algorithms: The p-Optimality Criteria // IEEE Transactions on Evolutionary Computation. — 2014. — V.18. — N 2. — P. 167–179.
10. Das. S. Differential Evolution: A Survey of the State-of-the-Art. // Das. S., Suganthan. P.N. IEEE Transactions on Evolutionary Computation. — 2011. — v.15. — N 1. — P. 4-31.
11. Мусаев А.А. Эволюционно-статистический подход к самоорганизации прогностических моделей управления технологическими процессами. // Автоматизация в промышленности. – 2006. – Вып. 7. — С. 31-35.
12. Мусаев А.А. Алгоритмы Data Mining в задачах управления динамическими процессами // Труды СПИИРАН. – 2007. – Вып. 5. — С. 299-312.
13. Metropolis N., Ulam S. The Monte Carlo Method. J. Amer. statistical assoc. — 1949. — 44. — N 247. — Pp. 335-341.
14. Ермаков С. М. Метод Монте-Карло в вычислительной математике: вводный курс / СПб. : Невский Диалект. —М. : БИНОМ. Лаборатория знаний. — 2009 . – 192с.
15. Редько В.Г. Эволюционная кибернетика. / М.: Наука. — 2001. — 159 с.
16. Емельянов В.В., Курейчик В.М, Курейчик В.В. Теория и практика эволюционного моделирования. — М.: Физматлит. — 2003. — 432 с.
17. Гудман Э.Д. Эволюционные вычисления и генетические алгоритмы // Обозрение прикладной и промышленной математики. — 1996. — Т. 3. — Вып. 5. — 179с.
18. David E. Goldberg. Genetic algorithms in search, optimization, and machine learning. // Addison-Wesley Publishing Co. — 1989. — 432p.