Аннотация

Головинский П.А., Суровцев И.С.

Системный анализ

В настоящей книге содержится изложение методов исследования сложных систем. Приведены базовые понятия и динамические модели, статистические оценки, методы исследование операций и методы искусственного интеллекта, применяемые в системном анализе.

Используемые динамические модели, теория нечетких множеств и методы искусственного интеллекта позволяют объединять междисциплинарные знания и технологии для решения сложных проблем. Центральной проблемой является формирование и принятие решений при разработке и коммерциализации инноваций, и применение в решении этих сложных задач динамических моделей, оптимизационных и интеллектуальных методов.

Учебник написан на основе преподавания изложенного в нем материала в Воронежском государственном архитектурно-строительном университете.

Книга может быть рекомендована в качестве учебного пособия в университетском образовательном процессе по направлениям и специальностям подготовки "Инноватика", "Прикладная информатика", "Управление инновациями", "Управление качеством" и др., а также будет полезна специалистам, работающим в сфере системного анализа, управления инновациями и менеджмента организаций.

Введение

Главной целью системного анализа является предоставление инструментов, помогающих при принятии решений, как в частной жизни, так и касающихся социальных и технических проблем, с которыми приходится сталкиваться в краткосрочной, среднесрочной и долгосрочной перспективе. Удачная характеристика системного анализа дана профессором Л. Хордиком, бывшим директором Международного института системного анализа в Люксембурге. Наиболее острыми в настоящее время представляются социально-технологические проблемы, которые окажут влияние на развитие всего человечества, на все стороны его жизни и на облик нашей планеты в целом. Системный анализ необходим при решении таких проблем, как обеспечение энергией и пищей, развитие промышленности и сохранение человеческого здоровья, состояние воздуха и воды, изменения в биосфере. Особенностью всех этих задач является то, что они являются комплексными и включают много важных аспектов, пренебрежение даже одним из которых может обойтись очень дорого.

Системный анализ имеет множество источников и связанное с этим множество подходов и методов. Поэтому в определении области применимости и методов системного анализа сохраняются определенные трудности. Однако ключевые черты системного анализа можно сформулировать вполне определенно. При рассмотрении проблемы методами системного анализа необходимо привлечь все имеющиеся в наличии знания в данной области, а при недостатке такого знания развить новые методы и подходы. Далее необходимо

четко сформулировать поставленную цель, поскольку от этого будут зависеть возможные пути ее решения. При решении подобных задач всегда существует целый набор возможных методов анализа и путей решений. На основе имеющихся знаний необходимо выбрать наиболее адекватные инструменты анализа данной проблемы и, изучив возможные альтернативы, выбрать наилучшее решение. При этом необходимо учитывать, как неопределенных характер будущего, так и реальные возможности организационных систем, имеющихся в нашем распоряжении. Результаты системного анализа должны быть представлены в виде, позволяющем определить дальнейшие шаги лицу, принимающему решения. В процессе реализации принятых решений они должны сопровождаться постоянной системной аналитикой, позволяющей лучше понимать суть происходящих изменений и своевременно вносить в них корректировку. Важнейшим инструментом системного анализа является математическое и компьютерное моделирование, позволяющее получать ответы на многие вопросы, возникающие на каждой стадии работы.

Системный анализ не только показывает возможные ограничения при принятии решений, но также позволяет вести в них значительную долю объективности и учесть побочные эффекты, обеспечивая новый взгляд на имеющуюся проблему и выбор наилучшего решения. Системный анализ вносит важный вклад в решение важнейших мировых проблем в быстро развивающемся инновационном мире. Освоение ключевых понятий и методов системного анализа позволяет лучше ориентироваться в принимаемых решениях и реалистичнее оценивать их последствия.

Головинский П.А., Суровцев И.С.

Системный анализ

Оглавление

Часть 1. Основные понятия и динамические модели

Глава 1. Основные понятия системного анализа

1.1. Понятие системы. *Система* – это совокупность элементов, связанных между собой и согласованно действующих для достижения определенной цели. Обязательными составляющими системы являются элементы, взаимосвязи и назначение (цель).

Примеры систем: может служить футбольная команда, школа, завод, правительство, живое существо. В футбольной команде элементами являются игроки, тренер, врачи, менеджер и др. Взаимосвязи между элементами футбольной команды формально прописаны контрактом, но содержат также значительное число неформальных элементов, которые могут оказывать самое существенное воздействие на поведение команды. Целью футбольной команды является максимальное количество побед над соперниками.

*Свойства систем***:** динамическое поведение, стремление к цели, адаптация к внешним условиям, самосохранение, эволюционные изменения.

Для рассмотрения выбранного сложного объекта как системы нужно последовательно ответить на следующие вопросы:

1) Можно ли идентифицировать ее составные части?
2) Влияют ли части друг на друга?
3) Могут ли части совместно дать результат, отличный от простой суммы вкладов отдельных частей?
4) Достигается ли результат системой при изменении внешних условий?

Многие взаимосвязи определяются через материальные потоки и потоки информации.

Цели системы могут быть не заданы явно или неверно декларированы. Лучший способ установить реальную цель системы – наблюдение за ее поведением.

Если правительство декларирует стремление к защите окружающей среды и не выделяет на это денег, значит, оно преследует иные цели.

Важнейшая цель любой системы состоит в обеспечении продолжения собственного существования. Система как целое может стремиться к целям, отличным от целей, которые преследуют составляющие ее элементы (муравьи, пчелы, студенты университета и сотрудники). Цель системы во многом подчиняет поведение системы элементов и системы в целом.

1.2 Динамика систем. Для понимания поведения системы во времени (динамики) важно правильно учитывать роль *запасов* и *потоков*. Запасы (или уровни) могут быть материальными (руда, нефть, вода, люди) и нематериальными – в виде знаний, проектов, опыта.

Простейшим примером регулятора, в котором присутствуют потоки и запасы (уровни) является ванна.

Рис.1.1. Ванна как простейший регулятор уровня

Соотношение входных и выходных потоков определяет изменение уровня в системе. Таким образом, уровень можно поддерживать двумя способами – регулируя входные или выходные потоки. Если мы рассмотрим в этом свете величину потребления нефти, то резкое увеличение энергоэффективности эквивалентно увеличению добычи нефти. Изменение запасов и уровней требует времени действия потоков. Поэтому отклик запасов на потоки и их изменение носит характер запаздывания. Это ключ к поведению систем. Действительно, леса не вырастают за одну ночь. Науку и технику можно развить только за десятилетия вложений. Если не вкладывать деньги в инфраструктуру, то она будет ветшать многие годы и не развалится сразу, но столь же длительным будет ее восстановление.

Инерция в поведении систем может использоваться для достижения поставленной цели так же, как мастера восточного единоборства используют импульс противника, чтобы победить его.

Запасы позволяют в течение определенного времени поддерживать *входные* и *выходные потоки* независимыми, тем самым, обеспечивая стабильность. Люди тщательно следят за запасами (денег, воды, корма для скота, топлива и т.д.).

Системы влияют сами на себя за счет *механизма обратной связи* (петли обратной связи). Он обеспечивает балансировку и устойчивость системы (*отрицательная обратная связь*). *Положительная обратная связь*, усиливающая имеющиеся в системе тенденции, разрушает систему.

1.3. Системы разных типов. Системы с одним запасом. Рассмотрим систему с одним запасом и двумя балансирующими циклами обратной связи на примере *обогревателя с термостатом*. Если температура в помещении падает ниже целевой температуры, датчик температуры улавливает разницу и подает сигнал включить

нагревательный элемент для подогрева воздуха. При достижении температурой в помещении желаемого значения, термостат выключает нагревательный элемент. В данной системе имеются два балансирующих (уравновешивающих) цикла – поток тепла от нагревательного элемента в комнату и отток тепла из комнаты на улицу. При достаточной мощности нагревателя цикл подогрева и остывания будет периодически повторяться. Подобным же образом приходится поддерживать запас товаров на складе. Из-за инерции систем текущий запас приходится поддерживать несколько выше номинально необходимого. Аналогично этому лицо, принимающее решение на основе обратной связи, не может изменить текущее поведение системы, вызвавшей эту обратную связь; все принимаемые решения повлияют только на ее поведение в будущем. Ни один поток не может повлиять на другой поток мгновенно. У любого балансирующего цикла обратной связи имеется некоторая переломная точка, после которой один из циклов становится доминирующим. Так при большой утечке тепла мощности нагревателя не хватит для поддержания необходимой температуры, и баланс сместится к меньшему ее значению.

Другим типом систем являются системы с *одним усиливающим механизмом и одним механизмом обратной связи*. Так изменяется численность населения и величина промышленного капитала.

Сложное поведение систем часто объясняется сменой доминирующих циклов во времени. Ценность системного анализа состоит в том, что он дает набор возможных сценариев.

Рис. 1.3. Схема баланса капитала

Системы с одинаковой структурой прямых и обратных связей демонстрируют схожие типы поведения. Поддержание постоянной температуры в комнате или запаса товара на складе носят характер колебаний. На величину колебаний можно влиять, разделяя управление на серию более мелких воздействий.

Системы с двумя запасами. Ресурсы, которыми обладает система, бывают возобновляемыми и невозобновляемыми, но в любом случае *рост в системе не может быть бесконечным* (машин, денег, капитала, численности населения). К невозобновляемым источникам относится нефть. Добыча нефти связана с двумя ресурсами – нефтью и капиталом. На начальной стадии процесс идет с усилением: большая добыча нефти дает больший капитал, позволяющий добывать больше нефти. С истощением запасов затраты капитала на единицу продукции растут, и добыча падает еще до полного исчерпания нефти. Чем быстрее растет добыча, тем быстрее будет пройден максимум. Возобновляемые источники могут обеспечить стабильность. К их числу относятся биоресурсы. Для возобновляемых источников ограничением является скорость восстановления.

Глава 2. Эффективность систем и их устойчивость

2.1. Устойчивость. Устойчивость систем является одним из их основных свойств. Устойчивость – это способность восстанавливать

состояние после прекращения внешнего воздействия. В случае систем близким к этому является понятие приспособляемости (адаптации).

В неживых системах устойчивость связана с упругостью и равновесием. В противоположном направлении работают хрупкость и жесткость. С определенными оговорками эти понятия можно применять и к живым системам. В сложных системах устойчивость обеспечивается несколькими механизмами обратной связи. Эти обратные связи позволяют восстановить или даже построить заново сами циклы обратной связи. В ряде систем существуют также *сверхустойчивость*, когда обратные связи могут самонастраиваться, иметь намерения (подцели), обучаться и эволюционировать в еще более сложные структуры.

Пример: устойчивость человеческого организма – терморегуляция, залечивание ран, обучение и др. У способности системы к самовосстановлению и устойчивости всегда есть пределы. Еще один пример – экосистемы. *Устойчивые системы могут быть весьма динамичными*. Реагируя на внешние изменения, они меняют свою численность и поведение. Неизменные, постоянные во времени системы, напротив, могут быть весьма хрупкими. *Статическая устойчивость* – стремление к восстановлению состояния равновесия. *Динамическая устойчивость* – повторяемость динамики на больших временах. Динамическая устойчивость – неочевидное явление, поэтому ее часто пытаются подменить статической устойчивостью, что может разрушить всю структуру.

Искусственно создаваемые системы часто обладают недостаточной устойчивостью: разведенные породы производительны, но подвержены болезням. Потеря устойчивость может произойти внезапно при плавном изменении параметров (бифуркации).

Самоорганизация – способность системы упорядочивать собственную структуру. Простейший пример – формирование снежинок. Самоорганизация типична для живых организмов, начиная с развития организма из яйцеклетки. Способностью к самоорганизации часто жертвуют в пользу краткосрочного увеличения производительности и стабильности. Этим же оправдывают снижение генетического разнообразия сельскохозяйственных растений. Эти же мотивы встречаются при управлении людьми. В итоге происходит деградация и разрушение системы. Способность к самоорганизации порождает разнородность и непредсказуемость.

Базовые правила самоорганизации часто просты, но на их основе возникают новые структуры, которые могут видоизменяться и усложняться. Примером такой простой системы, в которой возникают регулярные структуры, представляющие ячейки Бенара. Они возникают в тонком слое нагреваемого снизу масла за счет самоорганизации процессов конвекции в шестиугольные структуры на поверхности. В результате самоорганизации часто возникает иерархическое соподчинение, *иерархия* (военные системы, экономические системы и др.). Иерархии резко упорядочивают системы, придают им устойчивость и сокращают оборот информации. *Подсистемы иерархии* относительно независимы. Исходная цель верхних уровней иерархии – помощь нижним уровням в достижении их целей. Если интересы подсистемы достигаются в ущерб интересам системы в целом, такое поведение называется *субоптимизацией*. Примеры иерархий, в которых может возникать субоптимизация, дают корпорации, вузы и др. Необходим баланс централизованного контроля и автономности подсистем для функционирования самоорганизации.

2.2. Неожиданное поведение систем.

2.3. Ключевые точки. Важнейший вопрос теории систем состоит в том, как надо изменять структуру системы, чтобы приблизить ее поведение к желаемому. Обычно чувствительные точки системы известны, но их реакция на воздействие в этих точках чаще всего не соответствует нашим ожиданиям ввиду сильной нелинейности систем. Так, если система находится в хроническом экономическом застое, то изменение многих из ее параметров не приводит к изменению экономической системы страны.

Глава 3. Инновационное управление системами с неограниченной конкуренцией

3.1. Конкурениця. Конкуренция двух видов является хорошо известной задачей, которая подробно изучена в модели Лотка-Вольтерра системы «хищник-жертва». В экономике подобная модель Гудвина описывает классовую борьбу рабочих и капиталистов в виде системы двух дифференциальных уравнений для доли затрат на оплату труда и коэффициента занятости. Известно как существование колебательных режимов в этих моделях, так и их структурная неустойчивость. В то же время представляет интерес исследование конкуренции некоторой совокупности элементов одного вида, для которых способность к конкуренции является не одинаковой, а как-либо распределена по множеству элементов. Мы будем иметь в виду, например, конкуренцию между фирмами малого бизнеса одного профиля или между несколькими государствами, но у модели могут быть и другие применения, которые мы обсудим позднее.

Рассмотрим некоторые возможные сценарии перераспределения ресурсов как в чисто распределительных системах, так и в развивающихся системах с асимметрией производства и потребления

и возможности изменять эти сценарии путем изменения параметров элементов системы. На основе этого можно будет сделать определенные заключения о возможностях управления ресурсами в системах из конкурирующих элементов.

3.2. Основная модель конкуренции. Занумеруем конкурирующие n элементов системы индексом i. Конкуренцию мы понимаем как борьбу за ресурсы того или иного вида: финансовые, сырьевые или какие-либо иные в зависимости от типов рассматриваемых систем. Пусть все элементы системы однотипны и все они конкурируют индивидуально за одни и те же ресурсы. Долю ресурсов, находящихся в распоряжении i-го элемента системы в момент времени t, мы обозначим $w_i(t)$, а всю совокупность таких параметров будем описывать в виде вектора-столбца $\mathbf{w} = {}^t(w_1, w_2, \text{K}\ w_n)$. Каждый из элементов системы обладает некоторой способностью захватывать ресурсы в единицу времени, которую мы будем характеризовать коэффициентом f_i, а также противоположной способностью терять ресурсы в единицу времени, которую мы определим как g_i. Тогда скорость изменения доли ресурсов у i-й системы можно записать в виде

$$\frac{dw_i}{dt} = w_i f_i \sum_{j \neq i} g_i w_j - w_i g_i \sum_{j \neq i} f_i w_j,\ i, j = 1,2,\text{K}\ n. \qquad (3.1)$$

В силу симметрии условие $j \neq i$ можно опустить, поскольку соответствующие члены, учитывающие формально взаимодействие элемента системы с самим собой сокращаются. С учетом этого уравнение (3.1) можно записать в более компактном виде

$$\frac{dw_i}{dt} = w_i f_i(w, g) - w_i g_i(w, f), \qquad (3.2)$$

где

$$(w, g) = \sum_{j \neq i} g_i w_j,$$
$$(w, f) = \sum_{j \neq i} f_i w_j \qquad (3.3)$$

- соответствующие скалярные произведения.

Записанная система уравнений характеризует только возможности элементов системы в борьбе за ресурсы, которые всегда каким-либо образом распределены, оставляя в стороне вопрос о происхождении самих ресурсов. Здесь не учитывается возможность появления нового ресурса в каком-то из элементов системы, не связанного с перераспределением ресурсов между элементами системы, а возникшего, например, в результате производственной деятельности, обнаружения нового месторождения полезных ископаемых или технологического открытия. Если же таковой новый ресурс, связанный с деятельностью элементов системы появляется, то в правую часть уравнения (3.2) следует добавить слагаемое $h_i w_i$, что приведет как к суммарному росту всех ресурсов, так и к изменению динамики перераспределения ресурсов. Добавленное слагаемое описывает как производство ресурсов при положительном знаке константы скорости h_i, так и их усиленное потребление без воспроизводства при отрицательном знаке этой константы. При $h_i \equiv 0$ полный ресурс остается неизменным с течением времени. Действительно, просуммируем обе части уравнения (3.2) по индексу i. Обозначая $W = \sum_i w_i$, получим

$$\frac{dW}{dt} = (w, f)(w, g) - (w, g)(w, f) \equiv 0, \qquad (3.4)$$

то есть можно считать

$$W = \text{const} = 1. \qquad (3.5)$$

При $h_i \neq 0$ скорость изменения совокупного ресурса составит

$$\frac{dW}{dt} = (w, h). \qquad (3.6)$$

Отметим, что возможны как системы, развивающиеся с расширенным воспроизводством ресурсов, так и деградирующие системы с уменьшающимися ресурсами, а также ансамбли, состоящие из смеси разных элементов.

По всей видимости, почти распределительные системы тоже вполне возможны, например, в том случае, когда источником богатства является природная рента. При этом речь не идет, по сути, не о создании богатства, а его распределении.

Мы проследим возможную динамику перераспределения ресурсов в нескольких различных случаях, и начнем с распределительной модели, в которой $h_i = 0$.

3.3. Дискретная распределительная модель. При рассмотрении динамики системы особый интерес всегда представляет анализ возможности существования в ней устойчивых состояний. Предположим, что для системы (3.2) такое состояние возможно. Тогда $dw_i / dt = 0$, и

$$\frac{f_i}{g_i} = \frac{(w, f)}{(w, g)}, \forall i. \qquad (3.7)$$

Поскольку правая часть равенства (3.7) не зависит от индекса i, то $f_i = Cg_i$, т.е. способность элемента системы захватывать ресурсы пропорциональна его способности их терять, что представляется маловероятным, уникальным случаем. В силу этого, в общем случае устойчивое состояние в чисто распределительной модели невозможно.

Рассмотрим типичные особенности динамики распределительной модели для системы, имеющей всего два элемента, перераспределяющих между собой ресурсы. Система уравнений в этом случае имеет вид

$$\begin{aligned}\frac{dw_1}{dt} &= w_1 f_1 w_2 g_2 - w_1 g_1 w_2 f_2 = w_1 w_2 (f_1 g_2 - g_1 f_2),\\ \frac{dw_2}{dt} &= w_2 f_2 w_1 g_1 - w_2 g_2 w_1 f_1 = w_1 w_2 (f_2 g_1 - g_2 f_1).\end{aligned} \tag{3.8}$$

Обозначим константу $f_1 g_2 - f_2 g_1 = a$, тогда вместо (3.8) имеем систему уравнений

$$\begin{aligned}\frac{dw_1}{dt} &= a w_1 w_2,\\ \frac{dw_2}{dt} &= -a w_1 w_2.\end{aligned} \tag{3.9}$$

Пусть для определенности $a > 0$. В силу симметрии системы (3.9) это означает лишь последовательность нумерации элементов системы. Поскольку в силу нормировки $w_1 + w_2 = 1$, то

$$\frac{dw_1}{dt} = a w_1 (1 - w_1). \tag{3.10}$$

Начальное условие для уравнения (3.10) определяется начальным распределением ресурсов $w_1(0) = \nu_1$. Уравнение (3.10) является частным случаем уравнения Фишера-Прая, описывающего процесс вытеснения старой технологии новой. Решение уравнения (3.10) имеет простой вид

$$w_1(t) = 1 - \frac{1}{1 + A e^{at}}, A = \frac{1}{1 - \nu_1} - 1. \tag{3.11}$$

На больших временах $w_1(t) \to 1$, т.е. со временем ресурсы полностью переходят к первому элементу. Рис. 3.1 демонстрирует пример такого поведения.

Мы видим, что система, в которой имеется всего два элемента, неустойчива, что, в конечном счете, приводит к концент-

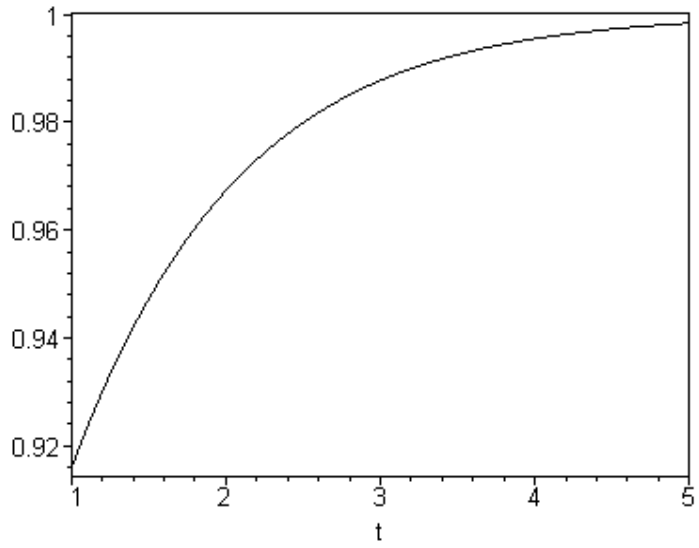


Рис. 3.1. Перераспределение ресурсов при чистой конкуренции двух элементов

рации всех ресурсов у наиболее сильного элемента независимо от начального распределения ресурсов.

3.4. Непрерывная модель. В данном разделе мы остановимся на ситуации, когда система содержит столь большое число однородных элементов, что распределение ресурсов по элементам можно считать непрерывным. В этом случае суммирование по индексу i можно заменить на интегрирование по некоторому непрерывному параметру τ. Разумно предположить, что достаточно общей является ситуация, когда элемент, имеющий большее значение коэффициента f_1 имеет меньшее значение коэффициента g_i и наоборот. Упорядочим ряд f_i в порядке возрастания. Тогда коэффициенты g_i с той же последовательностью нумерации образуют убывающую последовательность. Соответственно при переходе к непрерывным распределениям мы получим две монотонные функции: монотонно растущую функцию $f(\tau)$ и монотонно убывающую функцию $g(\tau)$. Конечно, такое формальное упорядочение может исказить смысловое содержание модели, расположив рядом достаточно разнородные в

других отношениях элементы. Однако данная модель различает элементы по их способности к конкуренции, поэтому такое упорядочение оправдано и резко сужает множество рассматриваемых функций.

Пусть текущий результат борьбы за ресурсы в сообществе описывается функцией распределения $W(t,\tau)$ по параметру τ, а t - текущее время. Функция распределения меняется с течением времени за счет взаимодействия элементов. С учетом парных взаимодействий кинетическое уравнение для функции распределения (3.2) приобретает вид

$$\frac{\partial W(t,\tau)}{\partial t} = f(\tau)W(t,\tau)\int W(t,\tau_1)g(\tau_1)d\tau_1 - \\ - f(\tau)W(t,\tau)\int W(t,\tau_1)g(\tau_1)d\tau_{.1}. \qquad (3.12)$$

Уравнение (3.12) является нелинейным интегро-дифференциальным кинетическим уравнением. Если проинтегрировать обе части уравнения (3.12) по значениям параметра τ, то получим

$$\frac{\partial \overline{W}}{\partial t} = 0, \qquad (3.13)$$

где $\overline{W} = \int W d\tau = \text{const}$. Это означает, что нормировка распределения остается постоянной во времени и ее можно принять равной единице. Таким образом, модель описывает перераспределение ресурсов внутри сообщества, в то время как полный ресурс остается неизменным.

Для удобства дальнейших вычислений сместим начало отсчета переменной τ, и изменим ее масштаб так, чтобы интервал ее изменений стал симметричным: $\tau \in [-1,1]$. Выберем в качестве функций, характеризующих интенсивность захвата и потерь ресурсов

линейные зависимости. Переопределим также масштаб времени так, чтобы учесть интенсивность процесса, определяемую постоянным множителем в правой части уравнения. Тогда вместо уравнения (3.12) получим уравнение

$$\frac{\partial W(t,\tau)}{\partial t} = W(t,\tau)[\tau - \int_{-1}^{1} W(t,\tau_1)\tau_1 d\tau_1]. \qquad (3.14)$$

Будем искать решение уравнения (3.14), удовлетворяющее начальному условию $W(0,\tau) = 1/2$, т.е. равномерному начальному распределению ресурсов. В силу положительности функции $W(t,\tau)$ возможна эквивалентная запись уравнения (3.14) в виде

$$\frac{\partial \ln W(t,\tau)}{\partial t} = \tau - \int_{-1}^{1} W(t,\tau_1)\tau_1 d\tau_1 . \qquad (3.15)$$

Правая часть уравнения (3.15) содержит два члена, один из которых зависит только от τ, а второй только от t. С учетом этого решение уравнения (3.15) имеет вид

$$\ln W(t,\tau) = \pi t + \ln F(t), \qquad (3.16)$$

где $F(t)$ неизвестная функция времени. Из условия нормировки функции распределения на единицу получим

$$W(t,\tau) = \frac{te^{\pi t}}{2\sinh t}. \qquad (3.17)$$

График решения представлен на рис. 3.2. Прямой подстановкой также можно убедиться, что получено верное решение.

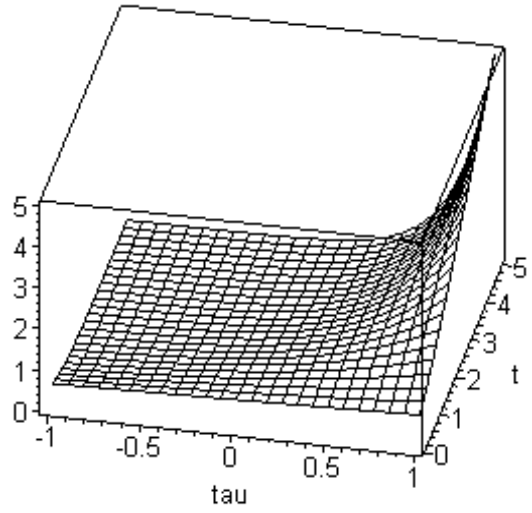


Рис. 3.2. Динамика перераспределения ресурсов с линейными конкурентными характеристиками

Из решения и его графического представления с очевидностью следует концентрация ресурсов с ростом времени в окрестности точки $\tau = 1$. Прямые численные расчеты подтверждают подобное поведение и для других конкретных зависимостей в функциях, описывающих конкуренцию.

3.5. Инновационные механизмы. Наличие подобной динамики означает вырождение любой системы с простой неограниченной конкуренцией с течением времени, когда «в живых» останется только один элемент системы. Скажем, с точки зрения конкуренции фирм, производящих однотипную продукцию это означает победу в конечном итоге одной сверхбольшой компании. Подобные процессы действительно постоянно происходят в мировой экономике, и для их корректировки в различных странах служит антимонопольное законодательство. Такой характер регулирования конкуренции является по сути внеэкономическим и, в определенной мере, силовым по отношению к самопроизвольным механизмам рынка. В связи с

этим представляют интерес модели с регулируемыми функциями конкуренции, а также модели, учитывающие дополнительные механизмы, отсутствующие в модели чистой конкуренции. Это позволяет ставить вопрос об управлении динамикой конкуренции.

Если законодательно менять время от времени условия, в которых происходит конкуренция, то такие меры будут препятствует накоплению ресурсов в одних руках. Поэтому так называемая «нестабильность законодательства», кроме известных негативных последствий, может нести в себе и функцию ограничения темпов роста концентрации капитала. Однако при этом все равно условия конкуренции остаются достаточно постоянными на протяжении длительных периодов времени, позволяющим в ряде случаев произойти процессам концентрации ресурсов. Другим способом поддержания конкурентной среды является поощрение постоянного возникновения новых конкурентно способных компаний. Возникновение таких компаний в той же среде, что и уже имеющиеся является практически бесперспективным, поскольку по описанным выше причинам они, скорее всего, исчезнут вскоре после своего появления. На самом деле этот процесс постоянно идет в мировой экономике в виде постоянного создания и достаточно быстрого исчезновения большого числа малых фирм, которые являются своеобразным «кормом» для крупных корпораций. Другой формой борьбы за существование на рынке является создание фирм, предлагающих товары и услуги отсутствующие в данный момент на рынке. Таким образом, речь идет об инновационном бизнесе, который не будет встречать конкуренции со стороны других производителей на начальном этапе своего развития, пока вновь не возникнет конкурентная среда. Отсюда видны преимущества инновационного механизма развития экономики с точки зрения поддержания

динамического равновесия в рыночной системе, поскольку он позволяет добиться хороших экономических результатов новым экономическим субъектам, в то время как борьба в имеющихся секторах для новых фирм обречена на поражение.

Результаты изучения модели перераспределения ресурсов при неограниченной конкуренции показывают, что система является абсолютно неустойчивой и в ней происходит неограниченная концентрация ресурсов. Найденное аналитическое решение в предельном случае модели с непрерывным распределением ресурсов по элементам системы показывает характерную динамику перераспределения ресурсов в больших системах. В приложении к конкуренции на рынке модель позволяет сделать вывод о монополизации рынка при отсутствии противодействующих механизмов. Проведенный анализ механизмов противодействия монополизации ресурсов показывает, что одним из эффективных методов поддержания динамического равновесия может служить политика инновационного протекционизма.

Часть 2. Статистические методы

Глава 1. Методы обработки, оценки и представления данных

1.1. Моделирование. Процесс управления и принятия решений включает в себя большое число детерминированных и случайных параметров и связей. Полное воспроизведение этих связей в какой–либо модели невозможно. Всякая модель представляет собой отображение реального объекта на иную систему связей, которая сохраняет наиболее важные характеристики изучаемой системы.

В зависимости от решаемой задачи более важными становятся те или иные характеристики и взаимосвязи в системе. Соответственно данная система может описываться целым набором различных моделей. Иначе говоря, выбор модели не является единственным. В то же время модель, построенная для описания какой-то конкретной ситуации, может иметь более общий характер и применяться для описания иных ситуаций и систем. Фактически развивается ограниченное число типов моделей, с помощью которых исследователи пытаются перекрыть все многообразие реальных ситуаций.

Математическое моделирование заключается в построении абстрактно-математического образа системы и протекающих в ней процессов. Чаще всего математические модели управления строятся для следующих целей:

- имитация действия объекта или протекания процесса при различных параметрах для получения представления об изменении при этом тех или иных характеристик системы;

- нахождение при помощи модели оптимальных параметров и режимов процесса;

- прогнозирование развития системы во времени с учетом детерминированных и случайных параметров.

Все эти задачи математического моделирования являются общими независимо от природы моделируемых объектов. Моделирование технико-экономических систем обладает своими особенностями, что и вызвало построение целого набора специальных моделей.

Для построения математической модели необходимо представить параметры и связи между объектами в математической форме. Для работы с количественными показателями их нужно, прежде всего, адекватно обрабатывать, то есть проводить статический анализ данных. Статический анализ данных с помощью современных компьютерных средств позволяет не только представить их в универсальной форме, но и выявить наличие или отсутствие взаимозависимости между ними.

1.2. Понятие о математической статистике. *Методы математической статистики* позволяют оценивать параметры системы и выделять в них случайную и закономерную составляющие. Хотя результаты наблюдений, зависящих от случайных факторов, нельзя достоверно предсказать, можно оценить шансы на их появление. Количественное описание шансов на тот или иной исход описывается на основе понятия *вероятности.*

Всё множество изучаемых элементов называется *генеральной совокупностью*. Если вся совокупность слишком велика, то приходится изучать часть этой совокупности. Такая группа элементов называется *выборкой* или *выборочной совокупностью*. Естественно выборку следует делать так, чтобы она наилучшим образом

представляла генеральную совокупность, то есть, как говорят, была *репрезентативной.* Наилучшей является случайная выборка, когда каждый объект имеет одинаковую вероятность быть выбранным. По выборке можно определить эмпирическую функцию распределения, оценить среднее значение выборки и дисперсию выборки.

Описательная статистика хорошо представлена в таких пакетах, как STATISTICA, MATLAB и STATGRAPHICS.

Важным средством анализа данных является графическое представление распределений – построение *гистограмм.* Стандартные графические пакеты позволяют также проводить проверку распределения на нормальность, то есть на соответствие гауссову нормальному распределению.

В теории вероятностей имеют дело со следующей ситуацией: если для некоторой системы выполняют ряд условий, которые называют *испытанием*, и в результате система переходит в некоторое определенное состояние, то говорят, что произошло *событие*. События можно разделить на следующие группы: *достоверное событие* – событие, которое обязательно происходит, если выполнено совокупность условий:

1. *невозможное событие* – событие, которое никогда не наступит при данных условиях;

2. *случайное* – наступление которого может произойти, а может не произойти.

Теория вероятностей не предсказывает появление того или иного случайного события, а определяет только *вероятность* его наступления. Случайные события в свою очередь можно разделить на:

1. *совместные,* когда два события могут наступить одновременно;
2. *несовместные*, когда наступление одного события исключает появление другого.

Если в результате испытания должно появиться одно и только одно из несовместных событий, то говорят, что события образуют *полную группу событий.* Классическое определение вероятности строится на основе понятия равной вероятности (равновозможности) событий. Равновероятными считаются события, если нет оснований считать, что частота наступления одного события в результате серии испытаний больше чем у другого.

1.3. Определение вероятности. Вероятность – это количественная характеристика шансов на исход исследуемого события. Вероятность события A обозначают как $P(A)$.

Пусть мы имеем дело с несовместными событиями, образующими полную группу. Назовем *элементарным исходом* каждый из возможных результатов испытания, а благоприятным исходом (события A) – исход, при котором наступает событие A. Тогда, *вероятностью события A* называют отношение числа исходов, благоприятствующих событию A, к полному количеству всех равновозможных несовместных элементарных исходов, образующих полную группу.

Иначе говоря, если событие A подразделяется на m частных случаев, входящих в полное множество из n попарно несовместимых и равновозможных событий, то вероятность $P(A)$ события A равна

$$P(A) = \frac{m}{n} \ . \qquad (1.1)$$

Вероятность $P(A)$ события A равняется отношению числа возможных результатов испытаний, благоприятствующих событию A, к числу всех возможных результатов испытаний. Вероятность удовлетворяет неравенству

$$0 \le P(A) \le 1. \qquad (1.2)$$

Вероятность наступления хотя бы одного из независимых событий равна сумме вероятностей отдельных событий, то есть

$$P(A_1+A_2) = P(A_1) + P(A_2). \qquad (1.3)$$

1.4. Условная вероятность. Вероятность события A при наложении дополнительного условия, что произошло событие B, называется условной. Рассчитаем вероятность. Пусть событию A благоприятствуют m событий, B – k событий, а пересечению AB – r событий (рис. 1.1.). Если произошло событие B, то это означает, что наступило одно из событий A, благоприятствующих B. При этом событию A благоприятствуют те и только те события, которые содержатся в AB. Таким образом

$$P(A/B) = \frac{r}{k}. \qquad (1.4)$$

Поскольку B по условию является достоверным событием, то число таких возможных событий k образует весь набор исходных возможностей. Таким образом, условная вероятность

$$P(A/B) = \frac{r}{k} = \frac{r/n}{k/n}. \qquad (1.5)$$

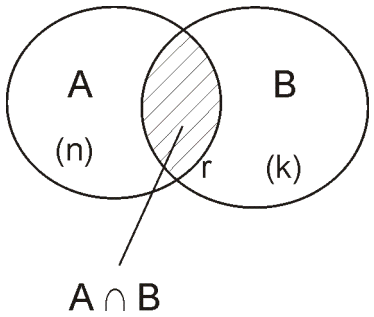


Рис. 1.1. Пересечение множеств событий

1.5. Случайные величины. Часто элементарному событию можно сопоставить число. Например, при случайном выборе человека из группы – возраст или вес человека, при случайном выборе частицы

газа – ее скорость, при случайном выборе момента времени диффундирующей частицы – ее энергию.

Функция, определенная на пространстве элементарных событий, называется *случайной величиной*. В случае дискретного пространства элементарных событий любую случайную величину можно задать, занумеровав в каком-либо порядке все точки пространства и связав с каждой из них соответствующее значение ***X***.

Пусть ***X*** – случайная величина, а x_1, x_2,... – ее значения. Совокупность всех элементарных событий, на которых ***X*** принимает определенное фиксированное значение x_i образует событие $\boldsymbol{X} = x_i$; его вероятность обозначается

$$P\{X = x_i\} = f(x_i),\ (i = 1,2,\mathrm{K}) \qquad (1.6)$$

и называется распределением вероятностей случайной величины ***X***. Ясно, что

$$f(x_i) \geq 0, \sum_i f(x_i) = 1. \qquad (1.7)$$

Аналогичным образом вводится распределение для двух и более случайных величин. Так для двух случайных величин событием считается $X=x_i$; $Y=y_k$, а функция

$$P(X = x_i, Y = y_k) = p(x_i, y_k),\ (i,k = 1,2,\mathrm{K}) \qquad (1.8)$$

называется *совместным распределением случайных величин* X и Y. Если случайная величина является непрерывной, то вместо функции распределения дискретной случайной величины вводят понятие *плотности вероятности*, а суммирование в (1.7) переходит в интегрирование.

Плотность распределения f непрерывной случайной величины x обладает свойствами

$$f(x) \geq 0,\ \int f(x)dx = 1. \qquad (1.9)$$

В природе и технике существуют самые разнообразные распределения, которые можно установить на основе проведения соответствующих полных или выборочных измерений.

Математическое ожидание M(X) называется также ожидаемым значением, средним значением или первым моментом:

$$M\left(X\right) \equiv \overline{x} = \sum_{n} x_k P\left(x_k\right). \qquad (1.10)$$

Для непрерывных величин

$$M(x) \equiv \overline{x} = \int x f\,(x) dx. \qquad (1.11)$$

Математическое ожидание величины x^r называется r-м центральным моментом x или r-м моментом x, а

$$D(X) = M\left(x - M\left(x\right)\right)^2 = M\left((x-\overline{x})^2\right) = \sigma^2. \qquad (1.12)$$

– дисперсия или второй центральный момент x, σ называется среднеквадратичным отклонением случайной величины от среднего значения.

При решении задач теории вероятностей важную роль играют *производящие функции* для вероятностей P_n, определяемые их разложением в ряд Тейлора

$$g(z) = \sum_{n} P_n z^n\,, \qquad (1.13)$$

где z – вспомогательная переменная и

$$P_n = \frac{1}{n!}\frac{d^n g}{dz^n}\bigg|_{z=0}\,. \qquad (1.14)$$

Производящую функцию (13) можно выразить через функцию случайной целочисленной переменной x в виде

$$g(z) = M(z^x). \qquad (1.15)$$

Действительно

$$M\left(z^x\right) = \sum_{n=0}^{\infty} P_n\left(x = n\right) z^n = g\left(z\right).$$

1.6. Нормальное распределение. Одним из наиболее распространенных является *гауссово нормальное распределение* (рис.1.2.)

$$f(x) = \frac{\exp(-x^2 / \ 2\sigma^2)}{\sigma\sqrt{2\pi}}. \qquad (1.16)$$

Практическую ценность нормального распределения определяет *центральная предельная теорема Ляпунова:* если случайная величина X представляет собой сумму очень большого числа взаимно независимых случайных величин, влияние каждого из которых на всю сумму пренебрежимо мало, то X имеет распределение, близкое к нормальному.

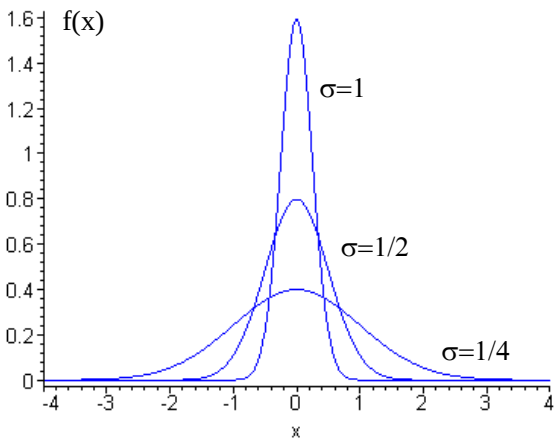


Рис.1.2. Вид нормального гауссового распределения

В приложениях часто приходится иметь дело со случайными величинами, которые слагаются из большого числа независимых компонентов. При очень широких условиях такие величины имеют нормальное распределение вероятностей.

1.7. Распределение χ^2 (Хи-квадрат). Если есть n нормально распределенных независисмых случайных величин χ, причем математическое ожидание каждой изних равно 0, а среднеквадратичное отклонение – единице, то сумма квадратов этих величин

$$\chi^2 = \sum_{i=1}^{n} \chi_i^2 \qquad (1.17)$$

распределена по закону χ^2 (хи-квадрат) с $k=n$ степенями свободы. Если эти величины связаны одним линейным соотношением, например $\sum_i x_i = n\bar{x}$, то число степеней свободы $k=n-1$. Плотность этого распределения

$$f(x) = \begin{cases} 0, \text{ где } x \le 0, \\ \dfrac{1}{2^{k/2}\Gamma(k/2)} e^{-x/2} \cdot x^{(k/2)-1}, \text{ где } x > 0, \end{cases}$$

$\Gamma(x) = \int_0^x t^{x-1} \mathrm{e}^{-t}\, dt$ – гамма функция, в частности $\Gamma(n+1) = n!$.

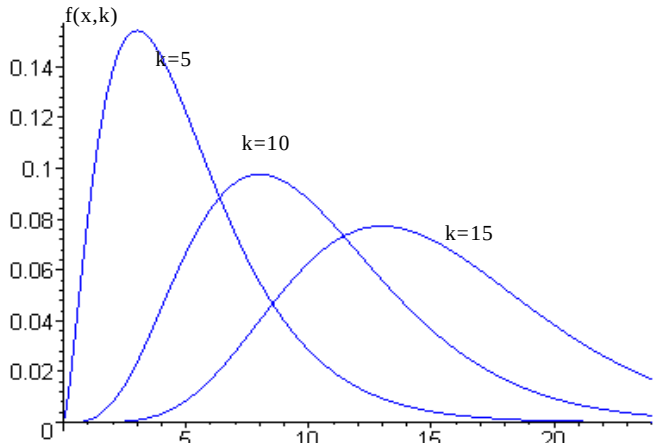


Рис.1.3. Распределение χ^2

Распределение χ^2 (рис. 1.3) определяется одним параметром – числом степеней свободы, с увеличением которого распределение плавно приближается к нормальному закону.

1.8. Корреляция. Две случайные величины могут быть независимыми или в той или иной мере зависеть друг от друга. Для характеристики меры зависимости двух случайных величин используется *коэффициент корреляции k(x,y)* случайных величин *X и Y*

$$K(X,Y)=\frac{M\left((x-\bar{x})(y-\bar{y})\right)}{\sigma_x\sigma_y}=\frac{M(XY)-M(X)M(Y)}{\sigma_x\sigma_y}=$$

$$=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (1.18)$$

где n – число пар точек $X=x_i$, $Y=y_i$. Используя свойство дисперсии

$$D(X)=M(X^2)-\left[M\left(X\right)\right]^2, \qquad (1.19)$$

формулу (17) можно представить в виде:

$$K(X,Y)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\cdot\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}. \qquad (1.20)$$

Для статистических независимых величин $K(X,Y)=0$. При согласованном изменении случайных величин $|K(X,Y)|=1$.

Глава 2. Линейный регрессионный анализ

2.1. Приближение табличных значений функций. При анализе тех или иных зависимостей мы зачастую не знаем закона, управляющего этой зависимостью. Обозначим эту неизвестную функциональную зависимость величины y от числовых переменных $x^{(1)}$, $x^{(2)}$,…$x^{(m)}$ как $y=f(x)$, и будем называть y откликом, а x –факторами, влияющими на отклик.

С геометрической точки зрения функция $f(x)$ задает поверхность в пространстве. Если $m=2$, то это поверхность в трехмерном пространстве ($x^{(1)}$, $x^{(2)}$, y) (рис. 2.1)

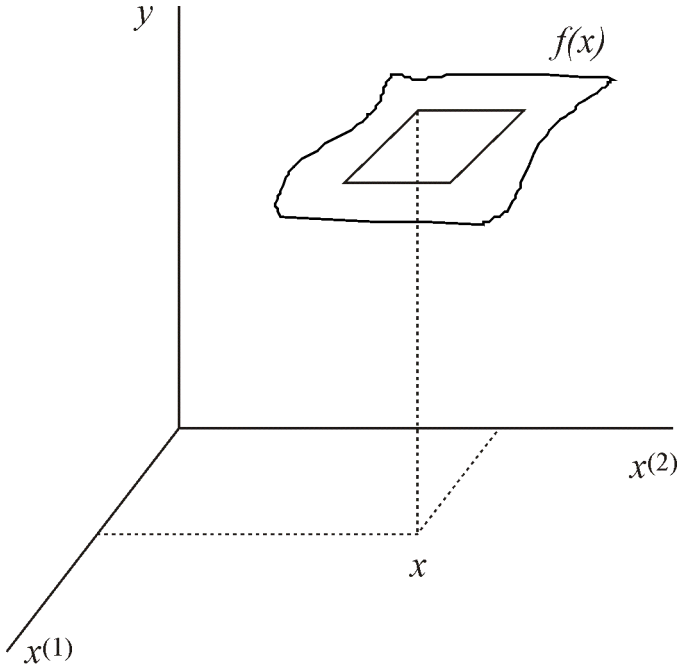


Рис. 2.1. Функциональная зависимость и ее линейное приближение

Простейшим подходом к определению $f(x)$ является замена неизвестной поверхности участком плоскости. Такая замена может хорошо описывать поведение функции в некоторой окрестности точки x:

$$y \approx \sum_{j=1}^{m} c_j x^{(j)} + c_0 + \varepsilon, \qquad (2.1)$$

где ε – погрешность.

Задача определения коэффициентов $c=\{c_i\}$, обеспечивающих наилучшее представление функции $f(x)$, заданной набором наблюдений y_k, $x_k^{(i)}$ (k=1,2,3…n), называется *множественной линейной регрессией*.

2.1. Нелинейная регрессия. Если подбираемая функциональная зависимость

$$y = F(x,c) + \varepsilon \qquad (2.2)$$

имеет более сложный вид, чем (2.1), то ее поиск путем определения коэффициентов c называется *нелинейной регрессией*. Задача нелинейной регрессии в общем случае является очень сложной. Самый простой случай регрессионных задач состоит в исследовании связи между одной независимой (одномерной) переменной x и одной зависимой переменной (откликом) y. Такая зависимость называется простой регрессией.

Исходными данными задачи регрессии являются два набора наблюдений $x_1, x_2, \ldots x_n$ – значения x и $y_1, y_2, \ldots y_n$ – значения y. Первым шагом в решении задачи является предположение в возможном виде функциональной связи между x и y. Для подбора вида зависимости y от x полезно построить и изучить график, на котором изображены точки с данными наблюдений (x_1, y_1), (x_2,y_2), $\ldots(x_n, y_n)$. Примерный вид зависимостей может быть также известен из теоретических соображений или предыдущих исследований аналогичных данных.

Часто используют следующие виды регрессий:

1) линейная - $y=a+bx$;
2) квадратичная - $y=a+bx+cx^2$;
3) степенная - $y=a_ix^i$;
4) логарифмическая - $y=a+b\ln(x)$;

2.3. Оценка точности регрессии. После подбора регрессионной модели следует выяснить, насколько хорошо выбранная модель описывает имеющиеся данные и найти наилучшее приближение. Единого правила для этого нет. Наиболее обоснованное решение можно принять путем сравнения

значений y_i со значениями, полученными с помощью регрессионной функции. Оценку точности аппроксимации нелинейной зависимостью можно оценить при помощи корреляционного соотношения

$$\eta = \sqrt{1 - \frac{\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}. \qquad (2.3)$$

Корреляционное отношение всегда удовлетворяет соотношению $0 \le \eta \le 1$. Если $\eta > K(X,Y)$, то кривая точнее аппроксимирует зависимость, чем прямая; для прямой $\eta = K(X,Y)$.

Разности ε_i между наблюдаемыми и предсказанными на основании регрессионной модели значениями называют *остатками.* Например, для линейной зависимости $y=a+bx$ значения остатков вычисляются как $r_i=y_i-f(y_i)=y_i-(a+bx_i)$, где a и b – оценки коэффициентов a и b.

При выборе методов определения параметров регрессионной модели можно использовать различные критерии, которые обеспечивают минимальность совокупных остатков. Можно, например, оценивать по сумме модулей остатков или по максимальному модулю остатка. Одной из наиболее удобных является оценка по сумме квадратов

$$\sum_{i=1}^{n}\left[y_i - f(x_i,c)\right]^2 \to \min. \qquad (2.4)$$

Оценка параметров c на основании критерия (2.4) называется *методом наименьших квадратов*. Метод наименьших квадратов широко используется в самых различных приложениях.

Рассмотрим применение метода наименьших квадратов для случая простой линейной регрессии $y = a+b(x - x)$. Для нахождения оценок a и b по методу наименьших квадратов нам надо выяснить, при каких a и b достигается минимум выражения

$$S = \sum_{i=1}^{n} \left[y_i - a - bx_i + b\bar{x} \right]^2 . \qquad (2.5)$$

Как известно, необходимым условием достижения минимума функции является равенство нулю ее частных производных

$$\frac{\partial S}{\partial a} = 0, \ \frac{\partial S}{\partial b} = 0 . \qquad (2.6)$$

Вычисляя производные, получим систему уравнений относительно *a* и *b*:

$$\sum_{i=1}^{n} \left[y_i - a - b(x_i + \bar{x}) \right] = 0 ,$$

$$\sum_{i=1}^{n} x_i \left[y_i - a - b(x_i + \bar{x}) \right] = 0 . \qquad (2.7)$$

Решение системы уравнений (2.7) легко найти: $a = \bar{y}$, где

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i , \qquad (2.8)$$

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} . \qquad (2.9)$$

В случае множественной линейной регрессии метод наименьших квадратов также приводит к системе линейных уравнений, которая легко решается на компьютере средствами линейной алгебры. В случае нелинейной регрессии получается система нелинейных уравнений, и поиск решения является сложной задачей. Не всегда ясно бывает и качество нелинейной регрессии.

Уверенность в том, что регрессия, даже линейная, правильно отражает опытные данные, никогда не бывает полной. Имеющийся опыт в этой области можно сформулировать в виде ряда требований. Первым и наиболее фундаментальным является предположение о том, что результаты отдельных

измерений представляют собой независимые случайные величины. Проверка этого свойства невозможна методами статистики, и оно обеспечивается всей методикой опыта. Вторым важным предположением является равномерность распределения ошибки наблюдения как случайной величины. Это означает, что измерения отклика имеют равную точность при всех значениях фактора, если случайную составляющую отклика рассматривать как ошибку при его измерении. В противном случае произвольной изменчивости ошибок классическая регрессионная схема непригодна.

Очень важным является предположение о виде функциональной зависимости, используемой для регрессии. Важно выбрать вид $f(x,c)$ так, чтобы она отражала определенные закономерности, хотя бывают эффективны и чисто эмпирические подгоночные формулы. Выбор регрессионной зависимости является одной из основных в любом исследовании.

Для возможности сделать на основании регрессии определенные статистические выводы, например, определить доверительный интервал для коэффициента регрессии, необходимо выполнение предположения о гауссовой статистике, то есть о распределении случайных величин по нормальному закону. Современные прикладные компьютерные пакеты позволяют легко проводить проверку на нормальность, например, распределения остатков r_i. Они также позволяют сделать заключение об адекватности регрессии.

Глава 3. Временные ряды

3.1. Характеристики временных рядов. При описании данных о развитии того или иного процесса во времени используется понятие временного ряда. *Временным рядом* называется случайная функция времени $x(t)$, заданная в определенные моменты времени t_1, t_2, t_3 … своими значениями x_1, x_2, x_3 …

Временные ряды могут быть *многомерными*, когда одновременно регистрируется несколько характеристик одного процесса в одинаковые

моменты времени. Временные ряды широко встречаются в практике. В экономике это динамика цен, изменение объема продаж, изменение производительности труда, изменение урожайности и т. д. Временные ряды можно представить графически. При этом они могут иметь самый разнообразный вид от четко выраженной периодичности (график продаж в течение суток) до случайных колебаний на фоне монотонного изменения величины (рост урожайности).

Анализ временных рядов представляет собой обширную область статистики. Временные ряды, рассматриваемые в разных предметных областях, имеют различные свойства. Поэтому для их исследования развиты и различные методы анализа. Мы затронем только ряд основных вопросов, относящихся к применению анализа временных рядов в задачах, связанных с принятием решений в технико-экономических задачах.

Анализ временных рядов преследует весьма широкие цели. В первую очередь он позволяет дать сжатое описание структуры ряда на основе подбора статической модели, описывающей ряд. Полученная модель позволяет предсказать будущие значения на основе предыдущих наблюдений, а также управлять процессом, порождающим временной ряд.

В реальности, однако, подобные цели далеко не всегда достижимы. Основным препятствием обычно является недостаточный объем статистических данных, и изменение статистической структуры временного ряда с течением времени. Эти изменения лишают ценности прошлые наблюдения, поскольку они уже не помогают делать прогноз динамики данных.

3.2. Анализ временных рядов. Анализ временных рядов проводится в несколько стадий:

1. Дается графическое представление временного ряда и проводится его визуальный анализ.

2. Из временного ряда удаляются закономерные составляющие зависимости ряда от времени – тренды и циклические (сезонные) составляющие.
3. Удаляются высокочастотные составляющие временного ряда (фильтрация).
4. Изучается случайная составляющая оставшегося временного ряда, и для нее подбирается математическая модель.
5. Осуществляется прогнозирование будущего развития процесса, представленного временным рядом.
6. Проводится исследование взаимозависимости между различными временными рядами

Все эти функции выполняются в современных статистических компьютерных пакетах.

При анализе временного ряда видимую его изменчивость стараются разделить на две составляющие – закономерную и случайную. Под закономерной (детерминированной) составляющей понимают последовательность значений d_i, вычисляемую по определенному правилу как функцию времени в момент t_i.

Наиболее часто используемые модели разложения временного ряда на детерминированную компоненту и случайную – это *аддитивная* и *мультипликативная модели*. В аддитивной модели временной ряд

$$x_i = d_i + \varepsilon_i, \qquad (3.1)$$

в мультипликативной модели

$$x_i = d_i \cdot \varepsilon_i, \qquad (3.2)$$

где ε_i – случайная компонента. Эти модели допускают переход от одной к другой, поскольку логарифмирование мультипликативной модели сводит ее к аддитивной:

$$\ln x_i = \ln d_i + \ln \varepsilon_i. \qquad (3.3)$$

Описание детерминированной компоненты временного ряда сильно зависит от области приложений.

В экономических задачах в *детерминированной компоненте* временного ряда d_i обычно выделяют три составляющие: тренд tr_i, сезонную компоненту s_i и циклическую компоненту c_i. В аддитивной модели

$$d_i = tr_i + s_i + c\,. \qquad (3.4)$$

В ряде случаев к этим компонентам добавляют компоненту, называемую *интервенцией*. Под интервенцией понимают крупномасштабно и кратковременное воздействие на временной ряд.

Трендом временного ряда tr_i называют плавно изменяющуюся непериодическую компоненту ряда. *Сезонная компонента* отражает присущую тому или иному процессу повторяемость во времени. Она часто присутствует в экономических, метеорологических и других рядах и состоит из последовательности почти повторяющихся циклов.

Типичным примером сезонного эффекта является увеличение объема продаж накануне праздников, увеличение объема пассажирских перевозок городским транспортом в утренние и вечерние часы. Главный смысл выделения сезонных компонент заключается в сравнении всех значений временного ряда через определенный период.

Наиболее часто используемые модели тренда:

1. линейная - $tr_i = a_0 + a_i i$, (3.5)
2. полиномиальная - $tr_i = a_0 + a_1 i + a_2 i^2 + \ldots + a_n i^n$, (3.6)
3. логарифмическая - $tr_i = \exp(a_0 + a_1 i)$ (3.7)

   (часто используется для описания процессов с постоянными темпами прироста),
4. логистическая - $tr_i = a\,(1 + a_1 \exp(-a_2 i))^{-1}$, (3.8)
5. Гомперца - $\log(tr_i) = a_0 - a_1 a_2^{\,i}$, $(0<a_2<1)$. (3.9)

Последние две модели задают кривые *s*-образной формы и соответствуют процессам с нарастающими темпами роста вначале и затухающими в конце процесса.

3.3. Анализ случайной компоненты ряда. Остановимся кратко на математических основаниях анализа случайной компоненты. Важным классом случайных процессов является *белый шум*, т.е. временной ряд с нулевым средним и независимыми составляющими его случайными величинами x_i при всех i. Частным случаем такой величины является *гауссовский белый шум*, представляющий собой последовательность независимых нормально распределенных случайных величин с нулевым средним и общей дисперсией σ^2.

Последовательности независимых случайных величин далеко не всегда являются адекватными моделями реальных временных рядов. Часто применяются *процессы скользящего среднего* (*moving average)*. Процессом скользящего среднего со средним μ называют процесс

$$x_i = \varepsilon_i + \theta\varepsilon_{i-1} + \mu, \qquad (3.10)$$

где $\mu = \bar{x} = \langle x \rangle$, θ – числовой коэффициент, i = 1,2,3,...n, ε_i – независимые случайные одинаково распределенные величины с нулевым средним значением. У процесса скользящего среднего статистически зависимы только соседние величины. Значения процесса с разностью $\Delta i > 2$ статистически независимы, поскольку в их формировании участвуют только разные статистически независимые слагаемые ε_i.

Часто в прикладных задачах встречаются *процессы авторегрессии (autoregression).* Процессом авторегрессии первого порядка (AR(1)) со средним значением μ называют процесс x_i, удовлетворяющий соотношению

$$x_i - \mu = \phi(x_i - \mu) + \varepsilon_i, \qquad (3.11)$$

где ϕ и μ – постоянные параметры. Если $\phi < 1$ то члены авторегрессии с ростом времени быстро становятся независимыми.

Если вероятностные свойства случайного процесса не меняются с течением времени, то он называется *стационарным случайным процессом*. Для оценки временных рядов широко используют математическое ожидание (среднее или первый момент)

$$\mu = \bar{x} = \langle x \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \tag{3.12}$$

и автокорреляционную функцию

$$r(k) = \mathrm{corr}(x_i, x_{i+k}), \tag{3.13}$$

где

$$r(k) = \frac{\sum_{i=1}^{n} (\bar{x}_0 - x_i)(\bar{x}_k - x_{i+k})}{\sqrt{\sum_{i=1}^{n} (\bar{x}_0 - x_i)^2 \sum_{i=1}^{n} (\bar{x}_k - x_{i+k})^2}}, \tag{3.14}$$

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{i+k} \ .$$

Величину *к* часто называют *задержкой* или *лагом.*

Для того, чтобы оценка (12) с ростом n приближалась к истинному значению (была *состоятельна*), достаточно, чтобы дисперсия

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (\bar{x} - x_i)^2 \to 0 \tag{3.15}$$

при $n \to \infty$. О стационарности процесса позволяет судить и анализ автокорреляционных функций процесса.

Сформулируем *теорему Слуцкого*: для стационарного процесса x_i оценка его среднего значения состоятельна тогда и только тогда, когда

$$\frac{1}{n}\sum_{k=1}^{m-1} r_i(k) \to 0 \tag{3.16}$$

при $m \to \infty$, где $r(k)$ – автокорреляционная функция процесса с лагом k.

3.4. Практический анализ и построение прогноза. При анализе временного ряда следует начинать с его графического представления, а затем выделить из него тренд. Процесс удаления тренда и других нестационарных компонент может проходить в несколько этапов. На каждом этапе анализируется ряд остатков, полученный в результате вычитания тренда из

исходного ряда. Признаком нестационарности ряда кроме графического анализа ряда служат не стремящиеся к нулю автокорреляционные функции и наличие ярко выраженных низкочастотных периодических пиков.

После максимального приближения ряда к стационарному виду подбирается статистическая модель объекта. Чаще всего используются параметрические модели авторегрессии – скользящего среднего (ARMA – модели). После подбора модели проводится оценка её адекватности и дисперсии остатков, которая далее используется для построения доверительных интервалов прогноза. Возможна ситуация, когда несколько моделей описывают временной ряд, так что результат неоднозначен. В ряде случаев модель подобрать не удается вообще.

Для выделения тренда чаще всего используется метод наименьших квадратов, который уже обсуждался нами в предыдущем параграфе. Часто нарушаются статистические предпосылки регрессионного анализа. При этом удается выделить плавные зависимости, но остатки обладают специфическими статистическими свойствами. Это не позволяет, например, оценить их дисперсию по остаточной сумме квадратов.

Наряду с методом наименьших квадратов для удаления тренда можно использовать и другие методы. Так линейный тренд может быть удален путем перехода к ряду разностей соседних членов ряда. Этот метод предложен Дж. Боксом и Г. Дженкинсом в 1970 г.

Процедура перехода от ряда x_i при $i=1,\dots n$ к ряду $y_i = x_i - x_{i-1} = \Delta x_i$ при $i = 2,\dots n$ называется взятием первых разностей. Оператор Δ называется простым разностным оператором первого порядка. Длина ряда y_i первых разностей на единицу меньше исходного ряда x_i .

Аналогичным образом можно ввести разностный оператор второго и более высоких порядков. Разностные операторы более высоких порядков позволяют выделять из ряда полиномиальные тренды соответствующего порядка.

После удаления тренда следует оценить сезонную компоненту по ряду $x_i - tr_i$. Если период p последовательности известен, то сезонную компоненту простейшим образом можно оценить как среднее

$$S_i = \frac{1}{m+1}\sum_{l=1}^{m+1}(x_{i+lp} - tr_{i+lp}) \quad i = 1,2,..p, \tag{3.17}$$

Усреднение производится по $m+1$ периоду, содержащему $n=(m+1)p$ значений ряда. Получив оценки сезонных компонент, их легко удалить из рассматриваемого ряда, путем вычитания из начальных значений ряда.

При наличии в ряде циклической компоненты часто для сглаживания ряда используется *метод скользящих средних*. В нем производится замена исходного ряда средними значениями в окрестности i на интервале времени, длина которого выбрана заранее. Таким образом, с изменением i среднее скользит вдоль ряда.

В качестве среднего часто выбирается среднее арифметическое

$$\overline{x}_i = \frac{1}{2m+1}(x_{i-m} + ... + x_i + x_{i+1} + ... + x_{i+m}). \tag{3.18}$$

Скользящее среднее сглаживает исходный ряд и даёт представление об общей тенденции – тренде и циклической компоненте. Чем больше выбирается интервал усреднения, тем более гладкий вид имеет график скользящих средних. Поэтому необдуманное применение процедуры со слишком большим интервалом сглаживания приведет к потере сезонной компоненты. С другой стороны, соседние члены ряда скользящих средних сильно коррелируют. Причина этого заключена в том, что в формировании ряда участвуют одни и те же члены исходного ряда. Это может приводить к тому, что ряд скользящих средних может порождать циклические компоненты, которые отсутствуют в исходном ряде (*эффект Слуцкого – Юла)*.

Метод оценки сезонных компонент при использовании метода скользящих средних в целом аналогичен процедуре с использованием

выделенного тренда. Если период последовательности известен, то для каждого сезона рассчитываются разности $x_i - \bar{x}_i$, а в качестве простейшей оценки берется простое среднее по m+1 периоду, т.е.

$$S_i = \frac{1}{m+1}\sum_{l=1}^{m+1}(x_{i+lp} - \bar{x}_{i+lp}),\ \ i = 1,\ldots p\,. \qquad (3.19)$$

В аддитивной модели дальнейшее удаление сезонной компоненты сводится к ее вычитанию из исходного ряда. В мультипликативной модели $x_i = c_i \cdot s_i \cdot r_i$ переходят к логарифмам, превращая модель в аддитивную.

Бокс и Кокс в 1964 году ввели семейство преобразований, позволяющих изменить масштаб данных временного ряда. Преобразование Бокса – Кокса применимо только к положительным числовым рядам и определяется с помощью формулы

$$F(x,\lambda) = \begin{cases} \dfrac{(x^{\lambda} - 1)}{\lambda}, \lambda > 0 \\ \log x, \quad \lambda = 0. \end{cases}. \qquad (3.20)$$

Преобразование Бокса – Кокса при $\lambda > 1$ растягивает расстояния между малыми значениями и сжимает его между большими по величине значениями. При $\lambda < 1$ наоборот, уменьшаются расстояния между меньшими значениями и увеличиваются между большими. Однако преобразование Бокса – Кокса существенно влияет на статистические свойства процесса и может значительно усложнить дальнейший подбор модели ряда.

Временные ряды можно анализировать с помощью пакетов SYATISTICA и STATGRAPHICS. Анализ временных рядов является специфической и обширной областью статистики, поэтому для их анализа используются также и специализированные пакеты.

Глава 4. Многомерный статистический анализ

4.1. Многомерные данные. Многомерные методы представляют графические и вычислительные средства для классификации и объединения элементов в группы на основе сходства и близости данных, представленных в виде множества переменных, относящимся к этим элементам.

Важной задачей многомерного анализа является снижение размерности, то есть выделение в пространстве параметров явления наиболее значимых координат. Одним из наиболее мощных методов многомерного анализа является кластерный анализ. Кластерный анализ позволяет построить дерево классификации n объектов посредством иерархического объединения их в группы или кластеры. Классификация строится на основании анализа расстояний в пространстве m переменных, описывающих объекты. В результате исходное множество объектов разбивается на подмножества компактных кластеров. Кластерный анализ не даёт оценки адекватности получаемых классификаций.

4.2. Метрика. Исходные данные представляются в виде матрицы размером $n \times m$. Пусть индекс i нумерует объекты, а индекс $к$ нумерует количественные признаки. Далее необходимо выбрать метод вычисления расстояния d_{ij} между объектами в многомерном пространстве. Вычисление расстояний между объектами с индексами i и j в зависимости от вида метрики производится по следующим формулам:

1. евклидова метрика

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}; \qquad (4.1)$$

2. сумма квадратов

$$d_{ij} = \frac{1}{n} \sum_{k=1}^{m} (x_{ik} - x_{jk})^2; \qquad (4.2)$$

3. метрика Манхеттен

$$d_{ij} = \frac{1}{n}\sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|; \qquad (4.3)$$

4. метрика Канберра

$$d_{ij} = \sum_{k=1}^{m}\frac{\left|x_{ik} - x_{jk}\right|}{x_{ik} - x_{jk}}; \qquad (4.4)$$

5. метрика матриц Брея – Кортиса

$$d_{ijj} = \frac{\sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|}{\sum_{k=1}^{n}x_{ik} + \sum_{k=1}^{n}x_{jk}}. \qquad (4.5)$$

Расстояние между данным кластером с объектами i и всеми другими кластерами j может вычисляться в соответствии со следующими стратегиями: ближайшего соседа $\min(d_{ij})$, дальнего соседа $\max(d_{ij})$ и ряда других.

Эффективные программы кластерного анализа содержатся в пакетах MATLAB и STATISTICA.

4.3. Факторный анализ. Переменные, описывающие исследуемый объект, могут иметь достаточно условный характер, лишь качественно отражая взаимозависимости и внутреннюю структуру объекта. Такие параметры называются *факторами*. Среди факторов можно выделять те, которые оказывают на некоторый показатель наибольшее влияние. Выделение таких наиболее существенных факторов и составляет предмет *факторного анализа.* Метод факторного анализа успешно применяется в нечетких науках - таких как психология, социология, экономика и др.

Факторный анализ включает несколько этапов. Вначале нужно выделить исходные факторы. Затем производится выделение *главных компонент*. Исходные данные как и в кластерном анализе представляются в виде значений m переменных для n объектов. Суть метода главных компонент состоит в следующем. Степень взаимозависимости переменных определяется уровнем коэффициента корреляции. В случае большого числа

переменных нужно по имеющимся данным вычислить матрицу R_{ij} корреляций между исходными переменными. Для определения наиболее коррелированных комбинаций факторов находят собственные значения матрицы R_{ij}. Располагая собственные векторы в порядке убывания, получают новые координаты в которых основных меньше, чем исходных. Это во многих случаях позволяет сократить необходимый набор переменных до двух – трех, что упрощает дальнейший анализ.

Исследование главных компонент наиболее плодотворно, когда все переменные имеют одинаковую природу и одинаковые единицы измерения, например, характеризуют структуру экономической деятельности предприятия (в процентах). Подчеркнем, что новые переменные, полученные путем диагонализации матрицы корреляции, могут не иметь прямого смысла и простой интерпретации. Однако, новые переменные указывают комбинации, в которых факторы максимально взаимодействуют.

4.4. Статистическое распознавание катастроф. Следуя В.И.Арнольду, будем называть *катастрофами* скачкообразные изменения, возникающие в виде внезапного ответа системы на плавное изменение внешних условий. Рассмотрим сборку Уитни, которая определяется решением уравнений

$$x^3 + ax + b = 0. \qquad (4.1)$$

Это означает, что при плавном изменении параметров a и b точка x будет находиться как корень уравнения (4.1). Графический образ уравнения показан на рис. 4.1.

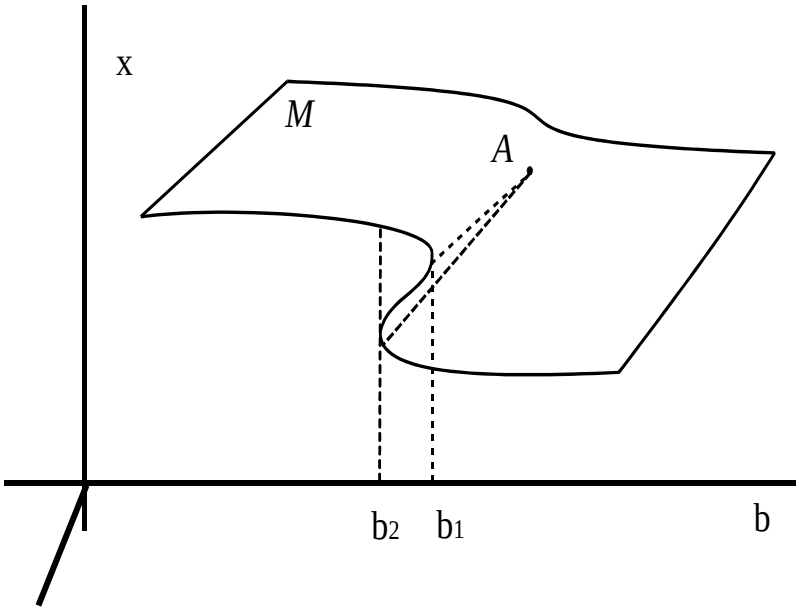


Рис. 4.1. Поверхность сборки

Из рисунка видно, что в области сборки положение точки на верхнем или нижнем листе поверхности зависит от маршрута движений в координатах (*a,b*). При одинаковых параметрах (*a,b*) в конечном состоянии значения *x* значительно отличаются. С ростом *b* точка скачком переходит на нижний лист.

Поставим вопрос о статистической различимости сборки Уитни в области А, где катастрофа только зарождается. Он связан с надеждой распознать катастрофу ещё на ранних подступах к ней. В области А ещё нет двузначности функций $x(a,b)$. В дальнейшем параметр *a* мы свяжем со временем *t*: $a \equiv t$. В реальной ситуации имеются временные ряды x_i, b_i, значения которых соответствуют изменяющемуся времени t_i. Задача заключается в попытке выяснить, не приближаемся ли мы к катастрофе, на основе анализа данных о временных рядах.

Распознавание катастрофы заключается в проверке того, насколько хорошо точки лежат на поверхности М. Функция $x(t,b)$ является двузначной,

в то время, как функция $b(x,\ t) = -x^3 - ax$ является однозначной. Такая неадекватность представления $x = x(t,b)$ позволяет рассчитывать на распознавание развивающейся катастрофы. Можно строить кубические аппроксимации

$$x = a_3 b^3 + a_2 b^2 + a_1 b + a_0 + d_1 t + d_2 t^2 \qquad (4.2)$$

и

$$b = -(c(x - x_0)^3 + t(x - x_0)) + b_0, \qquad (4.3)$$

а затем сравнивать их точность.

Если функция (4.3) значительно точнее аппроксимирует экспериментальные данные, чем (4.2), то далее ее можно исследовать на наличие максимумов и минимумов и определить наличие катастрофы. Обе зависимости (4.2) и (4.3) позволяют определить пары $(x_i,\ b_i)$ и $(x_i,\ b_i)$, которые несколько отличны в том смысле, что одна дает значения прямой функции, а другая – обратной. Поэтому их трудно сравнивать непосредственно. Для проведения такого сравнения определим $\tilde{x}_i$ по b_i на основании уравнения (4.2) и x_i' на основании решения уравнения (4.3), в котором выбирается корень s_i, наиболее близкий к x_i, то есть из условия $(|\ x_i - s_i|)$, где $i = 1$ иди 1,2,3 в зависимости от числа корней уравнения (4.2), которых может быть 1 или 3. Далее оценить близость к экспериментальным данным по величине

$$\tilde{S} = \sqrt{\sum_{i=1}^{n} (x_i - \tilde{x}_i)^2}\ , \qquad (4.4)$$

$$S' = \sqrt{\sum_{i=1}^{n} \left(x_i - x_i'\right)^2}\ . \qquad (4.5)$$

Сравнивая значения S и S' можно определить наличие катастрофы в структуре данных. Разумно сравнивать значения там, где есть три корня, то есть на интервале $[b_1,b_2]$, где b_1 соответствует минимуму функции (2), а b_2

соответствует её максимуму. На этом интервале следует оценить степень гистерезисности процесса. Для этого необходимо сконструировать безразмерный показатель $k = S'/S$. При $k \sim 1$ нельзя сказать ничего определенного. При $k > 1$ поверхность достаточно гладкая. При $k < 1$ мы имеем дело с потенциальной катастрофой. Желательно проследить эволюцию k со временем. Два возможных сценария показаны на рис. 4.2. Первый сценарий показывает, что мы имеем дело с ложной катастрофой. Второй сценарий явственно указывает на наличие катастрофы.

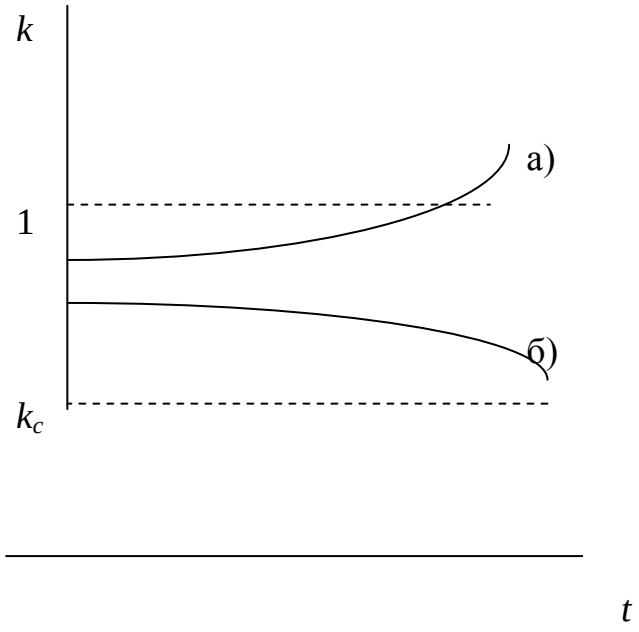


Рис. 4.2. Эволюция данных: а) – ложная катастрофа;
б) – катастрофа.

Локальный критерий катастрофы $k(t) < 1$, который является необходимым условием. Допустимый уровень $k(t)$, при котором можно с

уверенностью говорить о наличии катастрофы в структуре исходных данных, можно определить, видимо, только для конкретных задач.

5. Информационный анализ сложных систем

5.1. Информация в иерархических структурах. Понятие информации относится к числу основных представлений и не может быть определено исчерпывающим образом через другие понятия. Количественное измерение информации осуществляется с помощью известного определения Шеннона для количества информационной энтропии. Известно также, что уменьшение энтропии в системе с приростом информации о ней. В свою очередь изменение энтропии связано с термодинамической свободной энергией. Все эти связи хорошо известны, но их смысл по-прежнему широко обсуждается.

Определение количества информации можно произвести только относительно конкретной системы при заданной ее структуре. В этом смысле можно говорить об относительности информации и ситуация в определенной мере похожа на подход к проблеме наблюдателя в теории относительности или квантовой механике, где способ описания и конкретные параметры системы зависят от системы отсчета или макроскопического окружения системы. В квантовой теории поля параметры системы зависят как от выбора системы, так и от макроскопического окружения. Отметим, что между теорией относительности и квантовой механикой есть в этом отношении и принципиальное отличие. В теории относительности система отсчета может быть выбрана произвольно, в то время как поменять местами частицы и приборы в квантовой механике невозможно, - они входят в теорию несимметрично.

Вернемся к информации. Поскольку количество информации связано с некоторыми вероятностями, то в соответствии с основными положениями теории вероятностей, для определения этих величин необходимо определить

множество событий (множество элементов). При различном определении множеств по-разному будет определяться и вероятность, а, следовательно, и количество информации и энтропия. Таким образом, все эти характеристики относительны. При изменении определения структуры эти характеристики также изменятся. В случае изучения термодинамических свойств вещества минимальные структуры не вызывают сомнений, - это атомы и электроны. На разных уровнях иерархии имеются определенные структуры, которым можно приписать соответственно то или иное количество информации. Изменения, происходящие на разных иерархических уровнях, могут не влиять на количество информации друг друга. Например, нагрев или охлаждение газа в нескольких закрытых сосудах может не оказывать влияние на их взаимное расположение, в то время как информация, относящаяся к молекулам газа в этих сосудах, будет существенно меняться в этих процессах.

Если имеются иерархически независимые структурные уровни, то число возможных состояний всей системы является произведением состояний подсистем. Соответственно энтропия и информация для таких систем является суммой этих величин для отдельных подсистем. Изменение информации и передачу ее на разных структурных уровнях можно рассматривать как относительно независимые. Внешние воздействия на систему могут приводить как к непосредственному изменению макроструктуры, так и накоплению изменений на более простых структурных уровнях, которые достаточно долго не будут проявляться на макроскопическом уровне. Так нагревание раствора белка приведет по мере роста сначала к увеличению интенсивности хаотических перемещений молекул и их деформационных движений, а затем к коагуляции, то есть к резкому уменьшению соответствующей первоначальной структуре информации. Это пример того, что переходы с изменением информации не всегда обратимы. Последовательность обратимых переходы можно описать математически при помощи математической теории групп. В необратимых

термодинамических переходах количество информации уменьшается, а энтропия увеличивается. Последовательности состояний в таких процессах описываются полугруппами.

5.2. Принцип сохранения информации. Как известно, сложная устойчивая система, сохраняющая свою структуру, может быть только неравновесной. Поддержание такого состояния, то есть функционирование устойчивых неравновесных термодинамических структур возможно только для открытых систем. При этом выполняется принцип Николиса-Пригожина, согласно которому скорость производства энтропии в таком состоянии минимальна. Соответственно прирост информации в такой системе или потери информации о такой системе у внешнего мира также минимальны. В отсутствие потоков энергии и информации извне замкнутая система стремится к равновесию, увеличивая свою информационную энтропию и уменьшая долю сохраненной исходной информации.

Принцип Николиса-Пригожина можно принять как общий принцип функционирования сложных информационных систем в форме вариационного принципа для информации. Для биологических и социальных систем этот принцип нельзя доказать, и мы его будем использовать как постулат. Дальнейшее исследование подразумевает, в том числе, ответы на следующие два вопроса. Первый вопрос состоит в выяснении возможности получения уравнения, пригодного для применения к описанию динамики систем. С этой токи зрения представляет интерес согласованность этого принципа с непрерывной моделью эволюции. На этом пути требуется создать адекватный математический аппарат, без которого принцип останется в значительной степени описательным. Второй вопрос состоит в выяснении и обсуждении качественных следствий из принципа Николиса-Пригожина для различных сложных систем и их соотношения с наблюдаемыми фактами, то есть в его способности объяснить достаточно широкий круг явлений.

Изначально этот принцип в его общей форме относится к сложным системам, способным эффективно накапливать, хранить и перерабатывать

информацию. В естественной неживой природе такие системы неизвестны. Среди искусственных систем мы можем отметить компьютерные сети, в первую очередь Интернет, а в природе указанными свойствами обладают живые системы, начиная с простейших вирусов. Опыт изучения живых систем весьма велик. Основным принципом, объясняющим эволюцию живых организмов, остается теория Дарвина о происхождении видов путем естественного отбора. В соответствии с принципом естественного отбора малые мутации либо закрепляются, либо теряются в последующих поколениях живых организмов. Тем самым, если внешние воздействия, внешняя среда не меняются, то организмы, приспособившись к среде, могут далее длительное время существенно не меняться. Биологическая информация в живых системах сохраняется путем передачи ее потомкам с помощью механизмов размножения. Слабые мутации генетической информации обеспечивают приспособление организмов в цепи поколений за счет естественного отбора. Если поведение организма, основанное на всей имеющейся у него информации, не обеспечивает необходимый уровень адаптации, то организм гибнет, не давая потомства, и не воспроизводит тем самым свою биологическую информацию.

5.3. Сохранение информации в социальных системах. Подобные соображения в определенной мере применимы и к социальному поведению. Социальные институты, механизмы, ценности и традиции, обеспечивающие передачу информации следующим поколениям, выдержав влияние негативных внешних факторов, остаются устойчивыми, и информация, на которой они базируются, сохраняется. В случае, когда адаптация социальных структур к меняющейся внешней среде оказывается неадекватной в силу неадекватности применяемой информационной модели, погибают соответствующие социальные институты, а при отсутствии их адекватной для данной ситуации должной замены, погибает и целиком общество. В то же время удачная модификация информационной модели может способствовать повышению адаптивности системы. Такие изменения

информации закрепляются с помощью различных инструментов: законов, образования, науки, морали, религии и традиций. Перечисленные информационные инструменты приведены в порядке уменьшения их быстроты формирования и увеличения эффективности как регуляторов социальной среды.

В традиционных консервативных обществах, находящихся длительное время в изолированных квазистатических условиях, типичными являются неизменные и медленно меняющиеся социальные системы, а также застывшие языки. Например, в горных или удаленных районах целые народности остаются неизменными в течение столетий. В то же время наиболее интенсивно социальные институты меняются при воздействиях, которые коренным образом меняют условия существования общества. К таким воздействиям приводят, в частности, природные катастрофы и войны, а также вхождение в прямой контакт с социальными системами иной структуры.

Взаимодействие социальных структур с различной предысторией приводит к перемешиванию информации в этих системах или хотя бы в части из них. Если системы имеют совершенно разную историю и текущие условия существования, то можно ожидать, что предшествующая адаптация привела к накоплению в них принципиально разной информации, как описательного, так и процедурного характера. Приведем известный пример из биологии. Эволюция закрепила биологические молекулы с определенным типом симметрии относительно операции зеркального отражения (киральности). Симметричные им молекулы, например сахара, вполне можно синтезировать искусственно. Они будут иметь то же химическое строение, но не будут участвовать в обмене веществ живых организмов. Таким образом, изначально незначительная по величине информационная составляющая может с течением времени в процессе эволюции полностью изменять метаболизм организмов. Для социальных систем, существовавших по отдельности, накопленная в них информация также способствовала их

индивидуальной адаптации. Простое «перемешивание» информации способно привести к появлению сильно мутантной и, вполне вероятно, нежизнеспособной системы. В то же время возможны удачные гибриды, повышающие устойчивость системы и хранящейся в ней информации.

Подчеркнем, что последовательный расчет последствия значительного «скрещивания» информации в больших сложных системах не представляется возможным. Поэтому, во всяком случае, пока, невозможно заранее отделить его полезные последствия от вредных. Представляется более продуктивным понять основные закономерности таких процессов, моделировать их и вводить механизмы управления, улучшающие динамику информации, используя ее относительно медленное изменение на длительном интервале времени.

Некоторые механизмы, способствующие сохранению информации и ограничению темпов ее роста, использовались человечеством и ранее. Одним из таких механизмов является изоляционизм, характерный для определенных периодов исторического развития Китая, Японии и СССР. При этом происходило подавление информационного взаимодействия с внешним миром. Чаще всего такие меры нарушались извне путем вооруженного вторжения. В современном мире вооруженное вторжение обычно сопровождается, а иногда и заменяется информационным вторжением через средства массовой информации, а также путем распространения товаров и услуг, переносящих информацию. Даже естественнонаучная информация не является полностью инвариантной, поскольку формируется в определенном контексте общефилософских и общесоциальных представлений конкретного общества.

Практический вопрос, стоящий в связи с этим перед политиками, состоит в определении сфер и уровня перемешивания информации, способствующих максимальному сохранению информации того общества и той цивилизации, которую они представляют. Бесконтрольное перемешивание разнородной информации сильно изменяет количество и

содержание информации в данном обществе и может привести в последующем к полной потере исходной информации, например, в случае, когда механизмы копирования и размножения новой информации просто значительно превосходят по скорости механизмы копирования старой информации. Даже если поступающая в систему информация нейтральна по смыслу и не сказывается напрямую на механизмах функционировании системы, ее большое количество замедляет скорость работы с полезной информацией, а при превышении известного порога отношения сигнала к шуму фактически полностью парализует передачу информации в системе**.**

5.4. Перспективы технологического развития. Интересно рассмотреть некоторые тенденции технологического развития, которые следуют из принципа максимального сохранения информации. Прежде всего, стоит сделать определенные замечания относительно связи свободной энергии с энтропией. Здесь прямое соответствие имеется только на микроскопическом уровне, то есть на уровне микроинформации, в то время как на уровне макроинформации такой зависимости нет, хотя на создание информации всегда необходимо затратить некоторое количество энергии.

Если информационные процессы являются ключевыми в процессе эволюции живого вещества и человеческого общества, то естественная тенденция развития общества состоит в уменьшении количества свободной энергии, необходимой для создания, копирования или передачи единицы информации и увеличении величины свободной энергии, необходимой для уничтожения единицы полезной информации в системе. С точки зрения принципиальной направленности функционирования общества на обслуживание информационных процессов представляется естественным увеличение во всех технологических процессах информационной составляющей при снижении энергетических затрат. Это относится не только собственно к информационным технологиям, но и к функционированию машин и механизмов, а также к конструированию и производству все более «интеллектуальных» высокоструктурированных материалов.

Современные технологии расходуют свободную энергию весьма расточительно, но одновременно с этим имеющееся у нас знание демонстрирует возможность более эффективного ее использования. Следует ожидать, что зависимость технологий от энергии в среднесрочной перспективе, будет иметь тенденцию к радикальному снижению. Поэтому значительные энергетические проблемы, с которыми может в обозримом будущем сталкиваться человечество, носят скорее либо краткосрочный характер, либо характер отдаленной перспективы. В среднесрочном прогнозе нескольких десятилетий можно ожидать очень больших изменений в эффективности использования информации, что позволит приблизиться к теоретическому пределу, и приведет к быстрому снижению энергопотребления в развитых обществах. Высокое удельное потребление энергии грозит стать уделом слаборазвитых государств и обществ. Было бы ошибкой оставить без внимания возможность такого сценария развития технологий.

5.6. Взаимодействие социумов и конфликт цивилизаций. Рассмотрим некоторые проявления принципа сохранения информации применительно к взаимодействию социумов, то есть различных социальных структур одного иерархического уровня. Такие структуры, выполняя похожие функции, могут длительное время не приходить в соприкосновение, например в силу пространственной удаленности друг от друга, как исторически происходило с целыми цивилизациями. В современном мире географическая удаленность стран и народов играет все меньшую роль. Контакты цивилизаций на протяжении истории происходили чаще всего в форме острых конфликтов. Заслуга выявления конфликта цивилизаций как одной из стержневых сил развития современного мира принадлежит Тойнби, а осмысление современных политико-стратегических следствий из этого факта Хантингтону.

Очень важно, что Тойнби концентрирует внимание в проблеме конфликта цивилизаций не на борьбе за ресурсы или рынки сбыта, а на

конфликте культур, ценностей и религий, то есть на информационной компоненте взаимодействия цивилизаций. Анализируя проникновение информации в виде технологических и научных достижений, оба автора приходят к выводу, что частичное копирование больших массивов информации в контекст своей цивилизации сопровождается ее глобальным воздействием на всю систему информации цивилизации, для которой она является чужеродной, потенциально враждебной и губительной. При этом степень риска тем выше, чем больше масштабы вторжения такой информации и чем менее развиты механизмы переработки и защиты информации у данной цивилизации.

Принцип сохранения информации позволяет иначе взглянуть на природу ряда межличностных, расовых и этнических конфликтов. С позиции этого принципа они представляются не отдельными проявления дикости, нецивилизованности или следствием недостатков воспитания, а результатом действия фундаментального механизма, заложенного в живой природе и социальных системах. Данное положение совсем не означает принятие или морального оправдания таких явлений, но позволяет взглянуть на них с новой точки зрения, которая показывает с одной стороны ограниченность возможностей влияния на такие конфликты, с другой стороны указывает на возможные выборы инструментов воздействия на них. Иной взгляд возникает и на проблему глобализации, которая приводит к интенсивному перемешиванию информации в планетарном масштабе и фактически к сползанию в глобальную информационную войну в масштабах всей Земли. Фактически именно сейчас человечество вступило в ноосферную стадию развития, которую гениально предвидел Вернадский, однако реальность, как это часто бывает, оказалась куда более жестокой, чем представления великого гуманиста.

5.7. Образование и наука. Обобщенный принцип Николиса-Пригожина позволяет по-новому взглянуть на проблемы, связанные с образованием и наукой. С точки зрения принципа максимального сохранения

информации образование и наука выполняют важнейшие небиологические функции. По имеющимся оценкам количество информации, накопленное человечеством, не только многократно превосходит генетическую информацию человека, но и информацию, содержащуюся в структуре индивидуального мозга. Сопряжение внешней информации, накопленной обществом в самых разных формах, с индивидуальной информацией, содержащейся в мозге отдельного индивидуума, осуществляется с помощью различных способов обучения, совокупность которых мы понимаем как образование. Таким образом, образование напрямую связано с задачей сохранения информации в обществе.

Снижение уровня образования приводит к омертвлению, потере уже имеющейся в обществе информации, в том числе из-за потери понимания ее смысла и ценности, что неизбежно сказывается на ухудшении механизмов хранения и воспроизводства информации. Наиболее устоявшимся и распространенным способом хранения информации продолжают оставаться книги, а их местом хранения библиотеки. Тем самым, текущее состояние образования в стране проще всего в настоящее время оценить по состоянию ее библиотек. В России достаточно посетить Главную научно-техническую библиотеку в Москве, в которой нет ни читателей, ни современных журналов, чтобы убедиться, что образование находится в упадке. Эту горькую правду невозможно прикрыть дежурными комплиментами «высокому уровню образования» в нашей стране.

Все большую конкуренцию традиционным способам копирования и хранения информации на основе книгопечатания составляют компьютерные методы и системы. По всей видимости, в ближайшее десятилетие они сравняются с печатным словом по интегральной эффективности, а далее будет происходить быстрое вытеснение книг электронной информатикой. Пример подобного процесса наблюдается в цифровой фотографии, которая все больше заменяет традиционную пленочную технологию. С точки зрения перспектив развития России, есть смысл сконцентрировать усилия именно на

этой перспективной технологии с обеспечением доступа всего общества ко всем необходимым, в том числе платным ресурсам.

Еще одним важнейшим информационным социальным институтом является наука. Замечательная ценность науки для человечества заключается не столько в удовлетворении присущей человеку любознательности, как это часто декларируется, и как это иногда выглядит внешне, а в том, что наука резко изменила суть эволюционного процесса, заменив, в значительной степени, естественный отбор живых особей естественным отбором научных представлений и моделей. Может показаться достаточно удивительным, что наука, сама породившая теорию естественного отбора Дарвина, в течение длительного времени была не в состоянии взглянуть на собственное развитие с эволюционных позиций. Это удалось сделать только в середине двадцатого столетия Т.Куну. Он сделал акцент в развитии науки не на ее революционных моментах, а на так называемой «нормальной науке», устойчивое состояние которой назвал парадигмой. Малые изменения, - мутации в функционировании нормальной науки, приводят, в конечном счете, к научным революциям, по такому же механизму, как биологические мутации генов приводят к образованию новых видов. В общем виде в применении к человеческому познанию эти представления развиты Поппером.

Следуя и здесь обобщенному принципу Николиса-Пригожина можно сделать вывод, что, как и любая другая сложная информационная система, наука в действительности стремится к минимальному изменению информации, а наблюдаемые в ней научные революции явление столь же вынужденное и то же по сути, что и возникновение новых видов в живой природе.

Смещение давления естественного отбора от биологического отбора видов к отбору научных теорий является фундаментальным достижением человечества в его общем развитии. Всякое снижение уровня науки и образования в обществе тем самым должно активизировать более древние, по

сути классические, дарвиновские механизмы естественного отбора, вновь энергично привнося его в биологическую популяцию «человека разумного». Именно этот процесс, по многим признакам, наблюдался в России первое десятилетие после распада СССР. Нам представляется, что стратегические решения, принимаемые в области образования и науки, необходимо согласовывать как с фундаментальным принципом минимального роста информации, так и с его следствиями. В частности, из него непосредственно следует, что никакие быстрые «прорывы» в научных исследованиях принципиально невозможны. Они возникают в результате функционирования «нормальной науки» закономерным, но недетерминированным образом. Планирование открытий бессмысленно, но обеспечение эффективного функционирования «нормальной науки» многократно увеличивает выживаемость, конкурентоспособность всего общества. К сожалению, этот путь не является простым, и его нельзя ускоренно пройти в результате нахождения удачных административных решений. Достаточно обратить внимание на печальный опыт Германии, которая только в настоящее время, хотя и не в полной мере восстановила уровень своей науки, утраченный в период господства фашизма и в результате поражения во Второй мировой войне. В то же время, он вполне преодолим в исторически короткие сроки при нормальном функционировании науки, для чего требуется, понимание социальной необходимости и тенденций этого процесса, а также его правовое, инфраструктурное и финансовое обеспечение.

Часть 3. Исследование операций

Глава 1. Общая характеристика методов исследования операций

1.1. Основные понятия исследования операций. Исследование операций – математическая дисциплина, занимающаяся построением анализа и применением математических моделей принятия оптимальных решений. Основной задачей исследования операций является выбор в заданном множестве элементов удовлетворяющих тем или иным критериям. При этом любой элемент множества называют допустимым решением, а выбранный элемент оптимальным решением. Исследование операций подразумевает определенную последовательность действий в принятии решений и их реализации.

Типичные задачи исследования операций:

1.2. Задача о составлении рациона. Пусть имеется 4 вида продуктов, из которых необходимо составить паек удовлетворяющий следующим требованиям:

1. в паек должны входить все виды продуктов;
2. содержание белков, жиров, углеводов в пайке должно быть не менее установленных норм B;
3. стоимость пайка не должна превосходить некоторой величины C;
4. вес пайка не должен превышать величины P;
5. паек должен быть минимального объема;
6. паек должен иметь максимальную калорийность.

$x_к$ - единицы измерения, k = 1,2,3,4.

1. $x_к > 0$,
2. $\sum_{j=1}^{4} a_j \cdot x_j \geq B$,
3. $\sum_{k=1}^{4} c_k \cdot x_k \leq C$,

4. $\sum_{k=1}^{4} p_k \cdot x_k \le P$.

Система линейных неравенств задает на плоскости многоугольник, а в пространстве *выпуклый многогранник*. Выпуклой называется область, которая полностью лежит по одну сторону любой касательной.

5. $\sum_{k=1}^{4} V_k \cdot x_k \to \min V$.

6. $\sum_{k=1}^{4} q_k \cdot x_k \to \max Q$.

Условия 5 и 6, задающие в общем случае требование экстремума, называются целевыми функциями.

1.3. Задача о быстродействии. Корабль движется по реке со скоростью $\mathbf{V}(t)$, где t - текущий момент времени относительно течения, скорость которого постоянна как по величине, так и по направлению. Найти программу управления рулями корабля, при которой он достигает заданной конечной точки из заданной начальной точки за min время при условии $\mathbf{V}(t) = \mathbf{V}_0$. Это оптимальная задача быстродействия, т.е. задача о достижении заданного состояния системы за min время. Даная задача является типичной задачей оптимального управления динамической системой. Все задачи исследования операций делятся на 2 типа: на детерминированные и стохастические задачи, где поведение системы описывается вероятностью или имеется неполная информация о состоянии системы.

В первом случае говорят о *параметрических задачах* исследования операций. Во втором случае в *стохастических задачах* находят решения соответствующие наиболее желательным значениям параметра независимо от того, как конкретно реализуется неопределенность.

1.4. Задача о выборе наилучшей стратегии. Пусть имеется 2 игрока, первый из которых может избрать m стратегий, а другой n стратегий. В зависимости от выбранных стратегий выигрыш составляет величину A_{ij}(

$1 \le i \le m,\ 1 \le j \le n$). Если $A_{ij} > 0$, то выиграл игрок 1. Если $A_{ij} < 0$, то выиграл игрок 2. Критерием для каждого игрока является max его выигрыша.

1.5. Транспортная задача. Пусть в пунктах $a_1,\ a_2,\ldots,a_n$ находятся склады, в которых хранятся товары в количествах $X_1,\ X_2,\ldots,X_n$ соответственно. В пунктах $b_1,\ b_2,\ldots,b_m$ находятся потребители, которым нужно поставить эти товары в количествах не менее $Y_1, Y_2,\ldots,\ Y_m$ соответственно. Обозначим через d_{ij} стоимость перевозки единицы груза между пунктами a_i и b_j.

Исследуем операцию перевозки потребителям товаров в количествах, достаточных для того, чтобы удовлетворить потребности потребителей. Обозначим через x_{ij} количество товара, перевозимого из пункта a_i в пункт b_j. Для удовлетворения запросов потребителей необходимо выполнение неравенств

$$\sum_i x_{ij} \ge Y_j \ . \qquad (1.1)$$

Со склада с номером i нельзя вывести больше имеющегося на нем запаса, то есть

$$\sum_j x_{ij} \le X_i \ . \qquad (1.2)$$

Удовлетворить условиям (2.2) и (2.3), т.е. составить план, обеспечивающий запросы потребителей можно бесчисленным множеством способов. Выбор одного из них может основываться на одной из возможных оценок решения. Одним из критериев может служить минимальность стоимости перевозок

$$f(x) = \sum_{ij} d_{ij} x_{ij} \to \min \ . \qquad (1.3)$$

1.6. Задача об использовании ресурсов. Предприятие имеет в своем распоряжении определенное количество ресурсов: сырье, оборудование и т.п. в количестве $b_1,\ b_2,\ldots,b_n$. Пусть a_{ij} – число единиц ресурса i, необходимое для производства товара $j (j=1,2,\ldots m)$. Известно, что доход, получаемый от

единицы товара j, есть c_j. Обозначим через x_j количество товара j. Тогда доход предприятия

$$S = \sum_{j=1}^{m} c_j x_j \ . \qquad (1.4)$$

Общее количество использованных ресурсов i равно

$$\sum_j a_{ij} x_j \le b_i \ , \qquad (1.5)$$

поскольку не превосходит запаса b_i. Математически задача о распределении ресурсов состоит в отыскании неизвестных x_j, удовлетворяющих условиям $x_j \ge 0$, условиям (6) и сообщающих максимальное значение функции S.

Задачи 1) и 2) являются типичными задачам *линейного программирования*. Линейное программирование – это математическая дисциплина, изучающая методы нахождения наименьшего (или наибольшего) значения линейной функции нескольких переменных, при условии, что последние удовлетворяют конечному числу линейных уравнений и неравенств.

5.7. Задача составления расписаний. Теория расписаний – это большое направление в дискретной математике и теории исследования операций. Оно решает задачи выбора очередности и выделения определенного объема ресурса, который должен расходоваться на выполнение работ.

Одна из основных задач этого класса состоит в нахождении такого распределения ресурса и такого назначения очередности работ, при которых совокупность работ, составляющих проект, будет выполнена за минимальное время. Итак, пусть имеется перечень работ $P_1, P_2, P_3, \ldots, P_n$, необходимых для выполнения проекта, и требуемый ресурс для его выполнения. Ресурс может иметь различную природу. Это могут быть люди, оборудование, сырье, деньги. Таким образом, говоря о требуемом ресурсе, мы имеем в виду некоторый параметр, дающий перечень объемов ресурса различной природы.

Кроме того, выполнение работ бывает обычно стеснено ограничениями. Их, как правило, удается разбить на две группы:

а) К первой группе относятся ограничения, описывающие взаимную зависимость работ. Это ограничения логического характера. Типичный пример – ограничения типа графа, когда выполнению работы номера i предшествует некоторая совокупность работ, без выполнения которых нельзя приступать к работе с номером i. Так крыша здания не может быть построена раньше стен, а стены раньше фундамента. Ограничения этого типа могут быть сформулированы на языке теории графов. Виды работ можно обозначить вершинами ориентированного графа. Тогда ребра графа будут показывать, какие работы и в какой последовательности следует выполнять.

б) Второй тип условий связан с объемом ресурса, который может быть выделен на реализацию проекта. Обозначим $\mathbf{V}(t)$ – вектор ресурса, который может быть выделен в некоторый период t (t – номер месяца, квартала или года). Через $u^i(t)$ – обозначим долю работы номера i, через $\mathbf{q}^i(u^i(t))$ – вектор требуемого для выполнения работы ресурса. Тогда ограничения принимают форму

$$\sum_i \mathbf{q}^i(u^i(t)) \le \mathbf{V}(t). \qquad (1.6)$$

Если векторы $\mathbf{V}(t)$ заданы, то план реализации проекта сводится к следующему: для каждого интервала времени t должен быть указан перечень работ и доля $u^i(t)$ этих работ, которую необходимо выполнять так, чтобы суммарное время выполнения проекта было минимальным. Задача построения расписания – это задача *дискретного программирования*, в котором функции изменяются скачками, а аргумент (время) является дискретным.

6.8. Постановка задач оптимизации. Если целевая функция или ограничения заданы случайными функциями, то такие задачи относятся к *стохастическому программированию*. Если оптимальное решение ищется

для ситуации, развивающейся во времени, то такая задача решается методами *динамического программирования*.

Таким образом, *исследование операций* - это класс методов, относящихся к выбору способа действия, варианта плана, то есть к принятию решений. В этом смысле термин «операция» означает любое целенаправленное действие.

Исследование операций является синтетической дисциплиной, в которой можно выделить три главные стадии.

1. *Построение математической модели.* На этом этапе строится математическое описание системы.
2. *Описание операции.* При этом цель действий формулируется в виде некоторой оптимизационной задачи

$$f(x) \to \max, \tag{1.7}$$

где x – элемент некоторого множества, определяемого природой модели. Поиск максимума осуществляется с учетом налагаемых природой объекта ограничений и дополнительных условий, а также возможных неопределенностей.

3. *Поиск решения оптимизационной задачи*. Строго говоря, только эта стадия исследования операции относится к самой математике.

Глава 2. Линейное программирование

2.1. Постановка задачи. При всем многообразии конкретного содержания задачи линейного программирования имеют общую математическую структуру. *Линейное программирование* изучает методы нахождения наименьшего (или наибольшего) значения линейной функции нескольких переменных при дополнительных условиях, имеющих вид линейных уравнений и неравенств.

Это означает, что дана система линейных уравнений

$$\left.\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \text{L} + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \text{L} + a_{2n}x_n = b_2 \\ \text{L} \\ a_{m1}x_1 + a_{m2}x_2 + \text{L} + a_{mn}x_n = b_m \end{array}\right\} \qquad (2.1)$$

и линейная функция

$$f = c_1x_1 + c_2x_2 + \text{L} + c_nx_n. \qquad (2.2)$$

Требуется найти такое неотрицательное решение

$$x_1 \geq 0,\ x_2 \geq 0,\ \text{K}\ \ x_n \geq 0, \qquad (2.3)$$

системы (2.1), при котором функция f принимает наименьшее значение.

В реальных задачах часто ограничения накладываются на линейную комбинацию переменных в виде:

$$a_1x_1 + a_2x_2 + \text{L}\ \ a_nx_n + b_n \geq 0, \qquad (2.4)$$

Вводя дополнительное неизвестное

$$x_{n+1} = a_1x_1 + a_2x_2 + \text{L}\ \ a_nx_n + b_n \geq 0$$

сводим неравенство (2.4) к совокупности линейного уравнения и условию $x_{n+1} \geq 0$, т.е. приводим задачу к стандартной форме (2.1), (2.2), (2.3).

Часто в задаче линейного программирования требуется отыскать не минимум, а максимум линейной функции *f*. Так как

$$\max f = -\min(-f) \qquad (2.5)$$

то одна задача сводится к другой заменой f на $-f$.

Любую задачу линейного программирования можно сформулировать так, что все ограничения будут иметь вид неравенств. Из равенств (1) можно выразить r неизвестных x_1 , x_2 , ..., x_r в виде

$$\begin{aligned} x_1 &= d_{1,r+1}x_{r+1} + \text{L} \ + d_{1,n}x_n + \beta_1, \\ x_2 &= d_{2,r+1}x_{r+1} + \text{L} \ + d_{2,n}x_n + \beta_2, \\ &\text{M} \\ x_r &= d_{r,r+1}x_{r+1} + \text{L} \ + d_{r,n}x_n + \beta_r, \end{aligned} \tag{2.6}$$

Подставляя эти выражения для x_1 , x_2 , ..., x_r в функцию f , мы выразим f также через x_{r+1}, ..., x_n:

$$f = \gamma_0 + \gamma_{r+1}x_{r+1} + \text{L} \ + \gamma_n x_n .$$

В силу неравенств (2.3) исходная задача может теперь быть сформулирована в следующем эквивалентном виде:

Дана система r неравенств

$$\begin{aligned} x_1 &= d_{1,r+1}x_{r+1} + \text{L} \ + d_{1,n}x_n + \beta_1 \ge 0, \\ x_2 &= d_{2,r+1}x_{r+1} + \text{L} \ + d_{2,n}x_n + \beta_2 \ge 0, \\ &\text{M} \\ x_r &= d_{r,r+1}x_{r+1} + \text{L} \ + d_{r,n}x_n + \beta_r \ge 0, \end{aligned} \tag{2.7}$$

с n-r неизвестными x_{r+1}, ..., x_n , а также линейная функция

$$f = \gamma_0 + \gamma_{r+1}x_{r+1} + \text{L} \ + \gamma_n x_n . \tag{2.8}$$

Требуется среди всех неотрицательных решений $x_{r+1} \ge 0$, ..., $x_n \ge 0$ этой системы найти минимизирующее функцию f.

Любое неотрицательное решение системы ограничений называется *допустимым*. Допустимое решение, дающее минимум функции *f*, называется *оптимальным*. Саму функцию *f* часто называют *линейной формой*.

Отметим, что задача о поиске экстремума линейной функции многих переменных (2.8) не может быть решена обычными методами математического анализа. Действительно, частные производные f равны коэффициентам при неизвестных, которые в нуль одновременно не обращаются. Это означает, что функция f не достигает экстремума во внутренних точках области, задаваемой системой неравенств (2.7). Таким образом, если минимум f существует, то он достигается на границе.

2.2. Геометрическая интерпретация. Задача линейного программирования имеют геометрическое истолкование. Наиболее просто его понять в случае системы линейных неравенств с двумя переменными *x* и *y*. Тогда система ограничений имеет вид

$$ax + bx + c \geq 0. \qquad (2.9)$$

Пусть для определенности, $b>0$ тогда

$$y \geq -\frac{c}{b} - \frac{a}{b}x. \qquad (2.10)$$

Случаю равенства в (2.10) соответствует уравнение прямой, показанной на рис. 2.1.

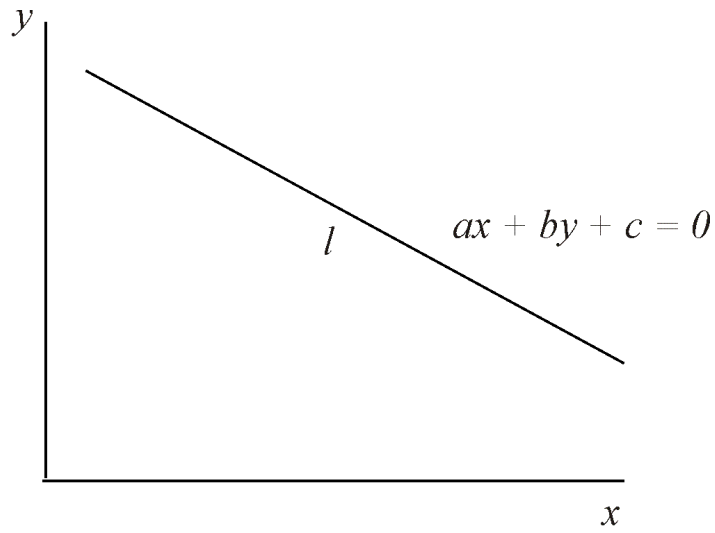


Рис. 2.1. Разбиение плоскости на две области неравенством (2.9)

Точки плоскости, удовлетворяющие неравенству (2.10), лежат выше прямой *l.*

Если имеется система неравенств

$$a_i x + b_i x + c_i \geq 0 \qquad i = 1,2,\text{K}\ n, \tag{2.11}$$

то она определяет пересечение системы полуплоскостей, представляющее собой некоторую многоугольную область *M,* как показано на рис. 2.1, 2.3.

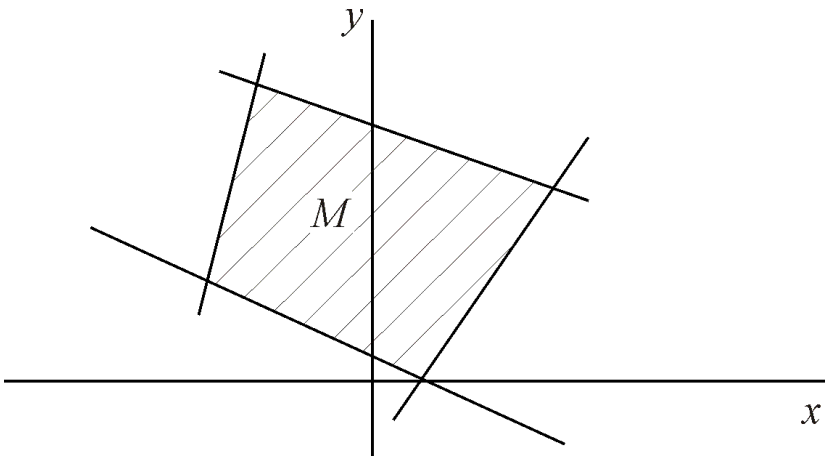


Рис. 2.2. Ограниченная область решений системы (2.11)

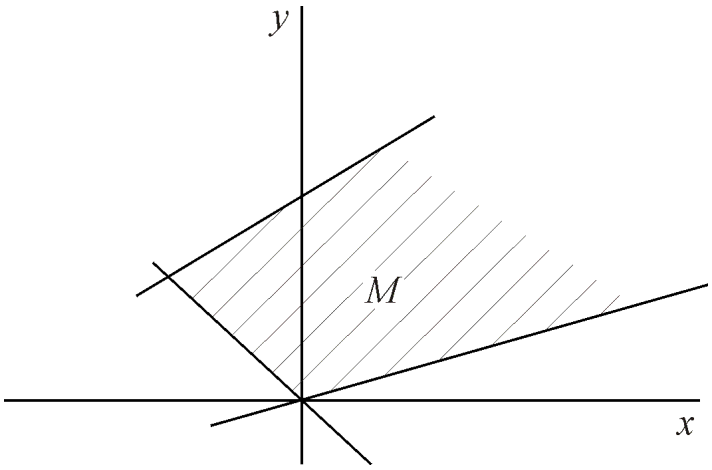


Рис. 2.3. Неограниченная область решений системы (11)

Область решений M является *выпуклой*. Это означает, что она целиком лежит по одну сторону от любого отрезка соединяющего две соседние вершины границы области M. Выпуклость многоугольника решений вытекает из того, что он получен путем пересечения нескольких полуплоскостей. В случае большего числа переменных область решений системы (2.7) образует выпуклый многогранник.

Выясним теперь геометрический смысл экстремума линейной формы f для двух переменных:

$$f = c_1 x + c_2 y . \qquad (2.12)$$

Зафиксируем временно некоторое значение f . Тогда этому значению соответствуют точки x и y, лежащие на прямой. Если значение f взять достаточно большим по модулю и отрицательным, то линия (2.12) будет лежать заведомо вне области M, если она замкнута, как показано на рис. 2.4.

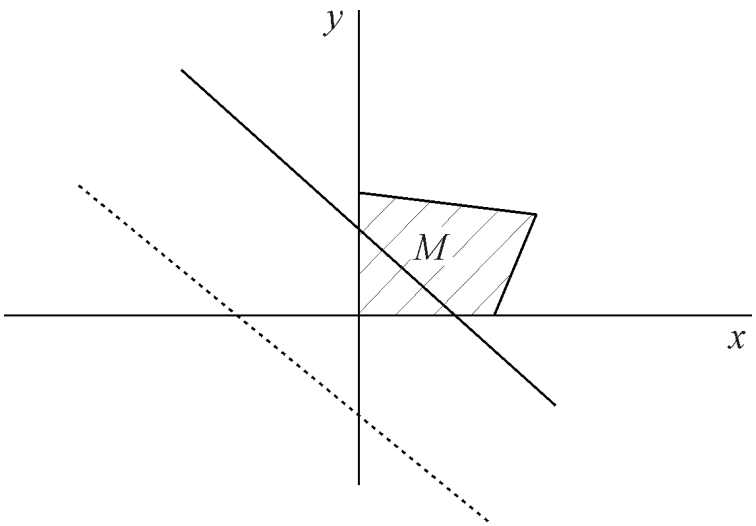


Рис. 2.4. Изменение взаимного расположения линии уровня и области решений при изменении f

Увеличивая значения f, мы будем смещать линии уровня параллельно друг другу, пока одна из них не коснется области решений. Точка касания линии уровня и области решений дает минимальное значение f в области M, то есть дает решение задачи линейного программирования. Возможен случай, когда встреча линии уровня с границей области M произойдет на отрезке ее границы. Тогда искомыми являются все точки этого отрезка.

Прямая, которая имеет с областью, по крайней мере, одну общую точку, притом так, что вся область лежит по одну сторону от этой прямой, называется *опорной* по отношению к этой области. С учетом этого исходная задача линейного программирования может быть сформулирована следующим образом: среди прямых уровня f найти опорную по отношению к области M, причем такую, чтобы вся область лежала со стороны больших значений f. Тогда любая из точек пересечения этой прямой с M дает решение задачи.

Геометрическая интерпретация задачи линейного программирования показывает, что ее можно решать последовательным переходом от одной вершины многогранника к другой с последовательным увеличением целевой функции f. Соответствующий эффективный алгоритм носит название *симплекс – метод* и реализован в виде компьютерных программ.

В ряде случаев переменные в задаче линейного программирования могут принимать только дискретные значения, как например построенные здания, количество машин, число бригад рабочих. В этом случае наиболее часто задача сначала заменяется непрерывной задачей линейного программирования, для которой существуют хорошо развитые методы. Дальнейший поиск целочисленного минимального решения проводится так называемым методом отсечения и заключается в нахождении целочисленной точки, которую первой пересекает линия уровня при изменении *f*, обеспечивающим ее движение от вершины многогранника вглубь области допустимых решений, как показано на рис. 2.5.

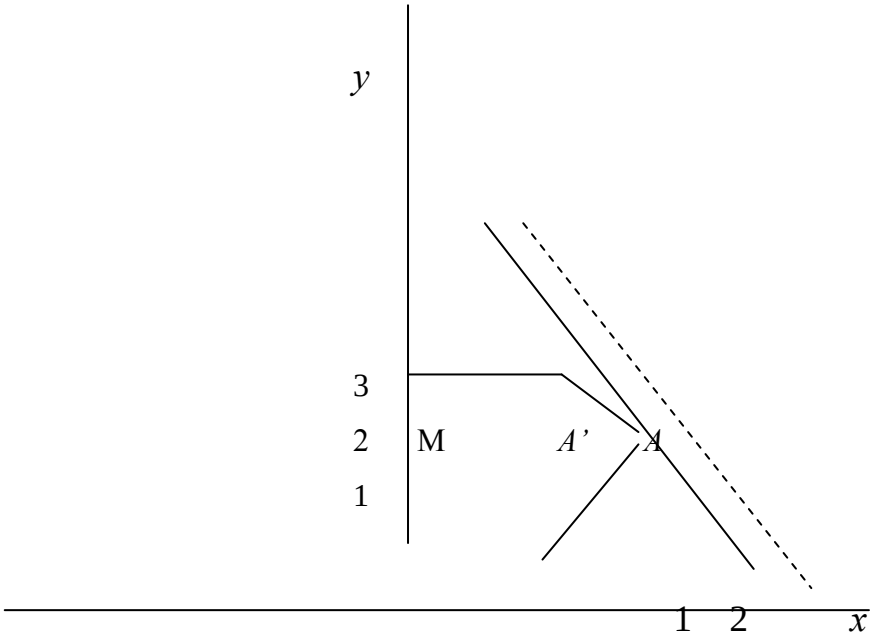


Рис. 2.5. Оптимальное решение для непрерывных переменных лежит в вершине A (пунктирная линия уровня), дискретное решение – точка A'

Глава 3. Сети и графы

3.1. Поиск на графах. Принятие решений эквивалентно поиску в некотором пространстве решений. При решении задач поиска возникают вопросы: «Гарантировано ли нахождение решений в процессе поиска, является ли поиск конечным или в нем возможно зацикливание (т.е. многогранное рассмотрение одних и тех же вариантов)? Является ли данное решение оптимальным? Как упростить поиск?»

Эффективным средством решения сложных задач поиска является представление задач в виде графов пространства состояний. Это позволяет использовать теорию графов для анализа структуры и сложности самой задачи и процедуры поиска ее решения. Граф состоит из множества вершин и дуг, соединяющих пары вершин. В модели пространства состояний решаемой задачи вершины графа представляют собой дискретные состояния процесса решений. Дуги графа описывают переходы между состояниями.

Теория графов является наилучшим инструментом исследования структуры объектов и их отношений. Сети с точки зрения их геометрии представляют собой узлы, соединенные связями-дугами, то есть графами. Возникает вопрос насколько такие структуры позволяют обеспечить представление и обработку данных, имеющих сложную иерархическую структуру, отображать различные связи и симметрии, способствовать представлению геометрических объектов разной степени сложности и реализовывать многовариантные стратегии поведения. Для ответа на эти и многие другие подобные вопросы необходимо иметь достаточно определенное представление об основных свойствах графов.

Отметим, что психолог Левин еще в 1936 г. высказал предположение о том, что «жизненное пространство» индивидуума можно представить в виде планарной (плоской) карты. На такой карте отдельные области представляют различные типы деятельности человека, например то, что он делает на работе, дома, или же его хобби (Lewin K. Principles of topological psychology. New York, 1939.).

3.2. Общие свойства графов. Существует два основных вида графов: неориентированные и ориентированные. *Граф* представляет собой множество точек, часть из которых или все соединены линиями. В *ориентированном графе* лини имеют направление от одной точки к другой. В *неориентированном* графе линии не имеют направления. Для графов не существенно расстояние между точками и форма соединяющих линий. Граф является дискретным объектом и может быть задан двумя дискретными множествами – множеством точек, которые называются *вершинами*, и множеством линий, соединяющих некоторые вершины - *ребрами*.

Любой граф можно разместить в трехмерном пространстве. Докажем это свойство. Разместим все вершины графа вдоль прямой линии. Выберем любую пару вершин графа. Их может соединять ребро. Если такое ребро есть, то проведем плоскость через прямую, на которой лежат вершины, и расположим ребро, соединяющие эти две вершины, в проведенной плоскости. Далее возьмем другую пару вершин, поместим их на прямую и повторим процедуру. В результате все выделенные ребра разместятся на веере плоскостей, пересекающихся по одной прямой.

Граф называется *простым*, если каждую пару вершин соединяет не более чем одно ребро. Граф называется *мультиграфом*, если хотя бы одну пару вершин соединяет более чем одно ребро. Ребра мультиграфа, соединяющие одну и ту же пару вершин, называются кратными. Примером мультиграфа могут служить два населенные пункта, соединенные несколькими дорогами. В простом графе ребро однозначно определяется парой вершин, которые оно соединяет, причем порядок вершин в паре не важен. В мультиграфе каждое ребро должно иметь свое собственное имя. Ребро, соединяющее вершину с самой собой, называется петлей. Граф, не имеющий ребер, называется пустым. Граф, в котором все вершины соединены между собой ребрами, называется полным.

3.3. Задание графа матрицами. Неориентированный граф задает два отношения между своими элементами: *отношение смежности* и *отношение*

инцидентности. Две вершины называются смежными, если они соединены ребром. Таким образом, *смежность* - это бинарное отношение, так как относится к двум элементам: V_i, $V_j \in V$, тогда (V_iV_j) - ребро. Вводят *матрицу смежности* следующим образом: C_{ij}=1 – вершины имеют общее ребро; C_{ij}=0 – вершины не соединены. Матрица смежности полностью задает граф. Пример графа показан на рис. 3.1.

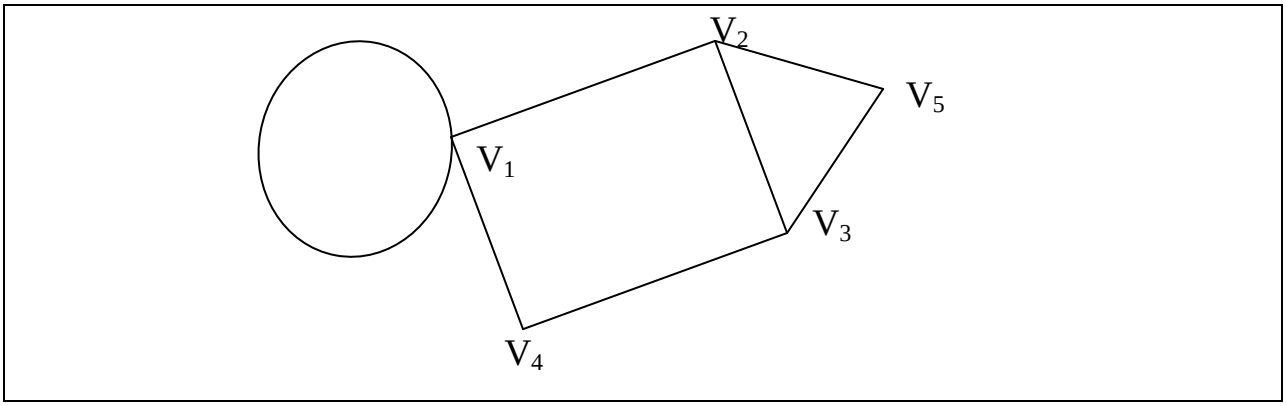


Рис. 3.1. Пример графа

Зададим его матрицу смежности:

$$C = \begin{pmatrix} 11010 \\ 10101 \\ 01011 \\ 10100 \\ 01100 \end{pmatrix}$$

Коэффициенты на диагонали соответствует петлям. Если по диагонали стоят нули, то значит петель нет. Для мультиграфа коэффициенты матрицы смежности равны числу ребер соединяющих вершины.

Одному и тому же графу могут соответствовать разные матрицы смежности, так как порядок нумерации вершин графа не важен. Таким образом, между матрицами смежности и графами нет взаимно-однозначного соответствия. Разным матрицам соответствуют разные графы, и, в то же время, одному графу могут соответствовать разные матрицы.

Инцидентность - это отношение между вершинами и ребрами. Ребро инцидентно каждой из вершин, которое оно соединяет. Инцидентность может быть задана прямоугольной бинарной матрицей *D*, в которой число строк равно числу вершин графа, а число столбцов - числу ребер: d_{ij}=1, если V_i инцидентно ребру l_j, d_{ij}=0, если V_i не инцидентно ребру l_j. Пример показан на рис. 3.2.

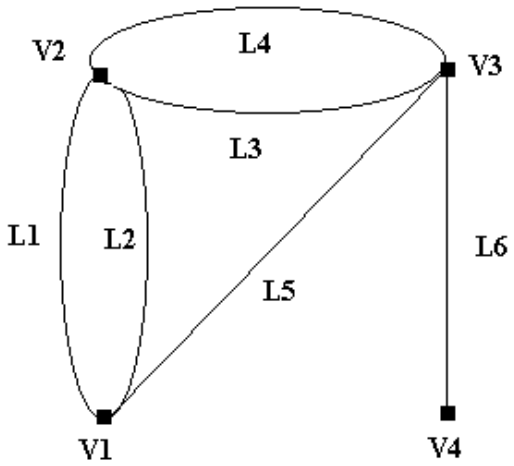


Рис. 3.2. Граф для построения матрицы инцидентности

Матрица инцидентности имеет для данного графа вид

	L_1	l_2	l_3	l_4	l_5	L_6
V_1	1	1	0	0	1	0
V_2	1	1	1	1	0	0
V_3	0	0	1	1	1	1
V_4	0	0	0	0	0	1

Таким образом, можно однозначно задать граф при помощи *матрицы инцидентности*. Число ребер инцидентных вершине V_i называется степенью

этой вершины. Вершина, степень которой равна единице, называется концевой или висячей. Граф называется подграфом некоторого исходного графа, если он содержит часть вершин исходного графа, соединённых также, как это было в исходном графе. Матрица инцидентности подграфа легко получается вычёркиванием соответствующих строк и столбцов. Между матрицей инцидентности и матрицей смежности существует алгебраическая связь.

3.4. Ориентированные графы (*орграфы*) задаются множеством вершин и множеством ориентированных рёбер. Ориентированное ребро графа соединяет вершину V_i с V_j и задаётся упорядоченной парой (V_i, V_j) с началом в V_i и концом в V_j. Ориентация ребра на графе указывается стрелкой. Примером орграфа может служить транспортная система с односторонним движением. Для орграфа его бинарная матрица смежности C в общем случае не симметрична. Коэффициент матрицы c_{ij}=1 только если $V_i \rightarrow V_j$. Если матрица симметрична, то это означает, что вершины соединяет не одно ребро вершину, а их минимум два и они направлены противоположно. Понятие инцидентности для орграфа сохраняется, но матрица инцидентности D различает начало и конец ребра: d_{ij}=-1, если j- начало ребра; d_{ij}=1, если j-конец ребра; d_{ij}=0, если нет соединяющего вершины ребра.

3.5. Пути и связность в графе. *Путь* в неориентированном графе – это последовательное соединение между собой первой вершины первого ребра в начале пути и последней вершины последнего ребра в конце пути. Число ребер в пути называется его длинной (L). При принятии решений длина – это число шагов с выбором, необходимых для принятия решения. Длина может принимать только целые значения. Путь называется циклическим или *циклом*, если начало и конец пути совпадают. Путь называется *цепью*, если каждое ребро в нем встречается не более одного раза, и простой цепью, если любая вершина графа встречается в нём не более чем 1 раз. Таким образом, простая цепь - это цепь, которая не пересекает сама себя.

Вершины V_i и V_j называются связанными, если существует путь с началом в V_i и концом в V_j. В этом случае говорят так же, что вершина V_j достижима из вершины V_i. Путь в ориентированном графе – это последовательность ориентированных ребер такая, что конец любого ребра совпадает с началом следующего ребра.

Имеет место следующая полезная **теорема для нахождения числа путей заданной длины**: Элемент $(i;j)=c_{ik}^{(L)}$ матрицы C^L, являющейся степенью матрицы смежности, равен числу путей длины L из $i \to j$. Доказательство производится по индукции, при этом для L=1 утверждение теоремы очевидно, поскольку по построению для матрицы смежности вершины i,j либо соединены напрямую и тогда L=1, либо не соединены, и тогда L=0.

3.6. Деревья. Неориентированный граф без циклов называется неориентированным деревом. Если граф без циклов несвязный, он называется лесом. Пример дерева показан на рис. 3.3.

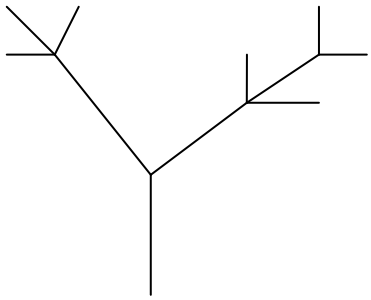


Рис. 3.3. Дерево

3.6. Планарный граф. Связный неориентированный граф называется планарным, если его можно уложить на плоскости. Граф укладывается на поверхности, если его можно изобразить на этой поверхности так, что никакие ребра не будут пересекаться. Плоским называется граф, который уже уложен на плоскости.

3.7. Стратегии поиска в пространстве состояний. Обычный поиск в пространстве состояний характеризует решение задачи как процесс нахождения пути решения. Под решением мы понимаем цепочку шагов, ведущих к решению задач от исходного состояния к целевому. Поиск в пространстве состояний можно вести в двух направлениях:

- от исходных данных задачи к цели;
- от цели к исходным данным.

При поиске на основе данных исследователь начинает процесс решения задачи, анализируя ее условие, а затем применяет допустимые ходы или правила изменения состояния. В процессе поиска правила применяются к фактам, которые в свою очередь используются для генерации новых фактов. Этот процесс продолжается, и в определенных случаях достигается цель.

Возможен альтернативный подход, когда движение в процессе решения ведется от цели, анализируются правила или допустимые ходы, ведущие к цели. Поиск от цели лежит в основе такого метода оптимизации как *динамическое программирование*. Определив направление поиска (от данных или от цели) алгоритм поиска должен определить порядок исследования дерева или более общего графа представляющего состояние и переходы в нижний уровень.

Альтернативными стратегиями являются поиск в глубину и поиск в ширину. Задача поиска в принципе может решаться простым перебором вариантов, однако практически это часто нереализуемо из-за большого числа возможных решений.

При поиске в глубину осуществляется последовательный переход от вершины к вершине при пошаговом удалении от начальной вершины графа. При этом в каждой вершине выбирается одна из ветвей с учетом имеющихся правил. Каждое следующее удаление на один шаг от вершины приводит в состояние, которое называется дочерним. Дочернее состояние генерируется правилами вывода, допустимыми ходами игры или другими операциями

перехода состояния. Если после исследования всего графа цель недостигнута, то поиск потерпел неудачу.

При поиске в ширину узлы графа рассматриваются по уровням. При этом исследуются те состояния, пути к которым короче. Тем самым при поиске в ширину найденное решение будет оптимальным в смысле длины пути (количество ребер) к целевой вершине.

Найденное решение при поиске в глубину может соответствовать не самому короткому пути. Существуют и разнообразные комбинации этих методов.

3.8. Эвристический поиск. *Эвристика* определяется как изучение методов и правил, открытий и изобретений. Эвристические методы ставят цели заменить полный перебор задачи поиска рассмотрением наиболее перспективных состояний.

Эвристика чаще всего используется в двух ситуациях:

1. Проблема может не иметь точного решения из-за неопределенности в постановке задачи или в исходных данных. Примерами могут служить диагностика и системы технического зрения, когда визуальная сцена зачастую неоднозначна, как это бывает при оптическом обмане.

2. Проблема может иметь точное решение, но стоимость его поиска может быть непомерно высокая.

Наиболее простой путь эвристического поиска – поиск ближайшего экстремума. Основой стратегии на поиске экстремума обычно оценивают не только текущее состояние поиска, но и его потомков. Для дальнейшего поиска выбирается наилучший потомок, при этом о его братьях и родителях просто забывают. Поиск прекращается, когда достигается состояние, которое лучше чем любой из его наследников. Такой поиск подобен тактики энергетического, но слепого альпиниста, поднимающегося по наиболее крутому склону до тех пор, пока он не сможет идти дальше. Так как в такой стратегии данные о предыдущих состояниях не сохраняются, то алгоритм не может быть продолжен путем частичного возврата из точки, которая привела

к неудачи. Ясно, что основной недостаток такого алгоритма – тенденция оставшаяся в локальном экстремуме, которая может не совпадать с глобальной.

Одним из самых распространенных алгоритмов является так называемый «жадный» эвристический алгоритм. “Жадный” алгоритм сохраняет списки пройденных состояний и возвращается после каждой пробы на шаг назад, чтобы в конечном итоге сделать шаг, наиболее приближающий к цели. Затем ситуация повторяется.

Глава 4. Оптимизационные задачи на графах и сетях

4.1. Порождающие деревья. Если граф конечный и связный, то легко построить дерево (и, как правило, не одно), множество вершин которого совпадало бы с множеством всех вершин заданного графа, а все ребра дерева одновременно были бы ребрами этого графа. Это можно сделать, например, так: пометим произвольную вершину графа, выберем какое-нибудь исходящее из неё ребро графа и пометим вершину, в которое это ребро входит. Если в графе есть ещё вершины, выберем ребро, выходящее из этих двух вершин в какую-нибудь третью вершину графа и пометим её. Если есть еще непомеченные вершины, то повторим процедуру, то есть выберем ребро, соединяющее одну из помеченных вершин с одной из непомеченных. Так как число вершин графа конечно (равно n), то потребуется (n-1) шаг для исчерпания всех вершин. Полученное таким образом дерево порождает граф или называется *порождающим деревом* графа или *остовом* графа, или его *стягивающим остовом*.

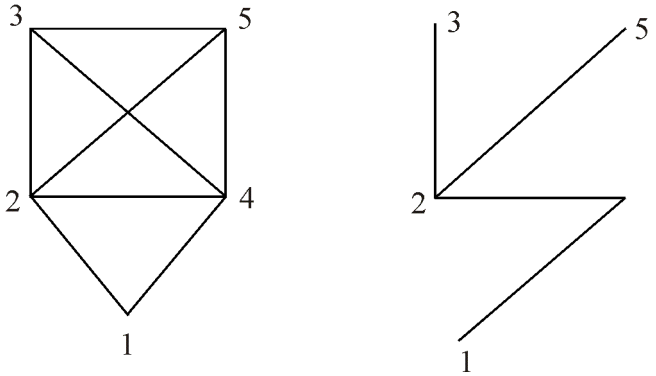


Рис. 4.1. Остов графа

Введем определение: множество всех ребер, выходящих из вершин графа, называется *звездой* этой вершины. Часто рассматривается задача, где каждому ребру графа приписано некоторое положительное число, которое называется его *весом*, тогда соответствующий граф называется *нагруженным графом* или *сетью.*

4.2. Задача о минимальном порождающем дереве. Среди порождающих деревьев сети имеется хотя бы одно, сумма длин всех дуг которого минимальна. Такое дерево называется *минимальным порождающим деревом* или *минимальным остовом.* Задача о нахождении минимального остова сводится к задаче о нахождении наиболее дешевой сети дорог, соединяющей некоторый набор населенных пунктов.

Пример 1: Пусть имеется некоторый набор населенных пунктов: A_1, A_2, A_3, …, A_7 , которые могут быть соединены так в виде графа, как показано на рис. 4.2. Требуется построить сеть дорог минимальной суммарной длины, то есть минимальное порождающее дерево.

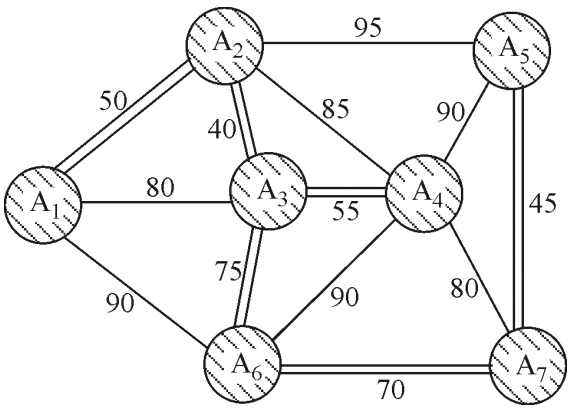


Рис. 4.2. Населенные пункты, соединенные минримальной по длине сетью дорог

Решение можно начать из любой вершины. Возьмем A_1; самый короткий путь (минимальный вес) A_1A_2. Рассмотрим 2 вершины и ищем минимальный вес, то есть A_2A_3. Далее процесс повторяется.

Алгоритм построения минимального остова.

ШАГ 1: Пометим произвольную вершину графа из ребер звезды, порожденной этой вершиной, выберем ребро минимальной длины, если имеются одинаковые минимальные – то любое из них, и пометим вершину, в которую входит это выбранное ребро. В результате две вершины графа оказываются помеченными. Если других вершин в графе нет, то искомое порождающее дерево построено и задача решена, иначе требуется сделать следующий шаг.

ШАГ 2: Каждая из двух помеченных вершин графа порождает свою звезду. Рассмотрим все ребра этих звезд, кроме ребер, которые соединяют уже помеченные вершины. Из этих ребер выберем наименьшее и пометим вершину, в которую это ребро входит. Если вершины исчерпаны, то построение завершено и задача решена, если есть еще вершины, то второй шаг нужно повторить, рассматривая звезды, которые порождают помеченные вершины. После (n-1) шага все n вершин графа окажутся помеченными, то есть задача будет решена.

Замечание: Шаг 1 удобно начинать с ребра наименьшей длины в сети. Результат применения алгоритма показан на рис. 6.2 двойными дугами, отмечающими сеть минимальной длины.

4.3. Задача о кратчайшем маршруте между выбранными вершинами. Пусть дана сеть и требуется найти кратчайший маршрут, ведущий из данного узла к каждому из других узлов в сети.

Пример 2. Рассмотрим работу алгоритма на конкретном примере из 7-ми населенных пунктов (рис 4.3).

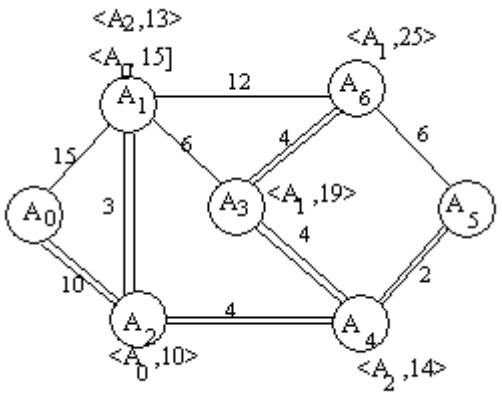


Рис. 4.3. Кратчайшая сеть для данного начального пункта

Требуется соединить A_0 со всеми пунктами кратчайшей сетью. Решение находится с помощью следующего алгоритма.

Шаг 1: дуга, связывающая A_0 и A_2, является самой короткой – это самый короткий путь в A_2. После окончания шага 1, узлы A_0 и A_2 имеют постоянные метки.

Шаг 2: отбирают все узлы, не имеющие постоянных меток, которые можно соединить с узлами, A_0 и A_2 одной дугой. Это узлы A_1 и A_4. Сравним длины маршрутов из A_0 в эти точки. Длина маршрута A_0 , A_1 составляет 15, длина маршрута A_0 A_2A_1 составляет 10+3=13. Так как 15>13 то временную метку $< A_0 ,15]$ заменяем на метку $< A_2 ,13]$, следовательно, узел A_4 получает метку $< A_2 ,14]$. Для узла A_1 маршрут A_0 A_2A_1 является кратчайшим, следовательно, выделяем дугу A_2A_1 , а метку $< A_2 ,13]$ делаем постоянной, то есть $< A_2 ,13>$.

Шаг 3: отбираем все узлы, которые соединены с узлом A_1 одной дугой и не имеют меток. Таких узлов два - A_3 и A_6. Узел A_3 получает временную метку - $< A_1 ,19]$, а A_6 получает метку $<A_1, 25]$. Среди узлов с временными метками снова выбираем узел с наименьшим расстоянием от узла A_0. Это узел A_4. Выделяем ребро. Далее процедуру повторяем. На 6-ом шаге получаем данный рисунок 4.3, где двойными линиями показан кратчайший маршрут.

Описание алгоритма. Сначала помечаем начальный узел и объявляем эту метку постоянной.

Шаг 1: Рассмотрим все дуги, идущие из начального узла, и припишем всем узлам, с которыми соединены эти дуги, временные метки. Временная метка состоит из элементов, с которыми соединяется узел и которые уже имеют постоянную метку, и числа, равного сумме длин дуг, соединяющих текущий узел с начальным через узлы с постоянными метками. Далее среди всех узлов с временными метками выбираем узел, расстояние от которого от начального узла минимально. Если таких узлов несколько – выбираем любой. Объявляем временную метку выбранного узла постоянной.

Шаг 2: Берем все узлы с постоянными метками, и рассмотрим все дуги идущие из этих узлов с постоянными метками. Вершины, в эти которые дуги приходят, мы пометим временными метками. Временные метки этих узлов содержат элемент «откуда» ребро пришло и расстояние от начального узла. В результате мы получим новый набор узлов с временными метками, образовавшийся на предыдущем шаге и на данном шаге. Все временные метки нужно сравнить и выбрать ту, которая имеет наименьшее расстояние до начального узла. Эту временную метку делаем постоянной и соединяем её постоянной дугой с узлом, указанным в метке. Дальше шаги повторяем до исчерпания сети. Данный алгоритм заканчивается за (n-1) шаг.

4.4. Задача о максимальном потоке. Сеть называется *ориентированной,* если ориентированы все ее дуги, то есть сеть – это нагруженный ориентированный граф.

Задача о потоках возникает при построении следующих моделей:

1. перемещения товаров и грузов;
2. перемещения денежных масс;
3. передачи данных компьютерными сетями;
4. водопроводных сетей;
5. электрических сетей;
6. транспортных потоков.

Узел сети, который является начальным для всех своих дуг (все эти дуги являются выходящими), называют *источником*. Узел, который является конечным для всех своих дуг (все дуги являются входящими), называется *стоком*. Все другие узлы такой сети называются *промежуточными*.

Припишем каждой дуге A_iA_k ориентированного графа число $c_{ik} = c(A_iA_k)$ – то есть пропускную способность этой дуги.

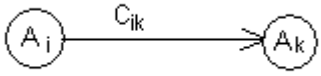


Рис. 4.4. Дуга ориентированного графа с указанием ее пропускной способности

Пропускной способностью называется максимально возможное количество продукта (например, жидкости или газа в трубопроводе), которое может быть доставлено из узла A_i в узел A_k за единицу времени. Такую сеть нередко называют *транспортной*.

В дальнейшем будем считать, что в транспортной сети есть ровно один источник и ровно один сток. Это ограничение не является существенным, так кака все источники можно объединить в один виртуальный источник, и также все стоки объединить в один виртуальный сток.

Рассмотрим транспортную сеть с одним источником A_0 и одним стоком A_n, а промежуточные узлы обозначим: A_1, A_2,…, A_{n-1}. Будем говорить, что в сети задан поток величины v, если каждой ориентированной дуге A_iA_k приписано неотрицательное число $\varphi_{ik} \geq 0$, которое называется потоком по дуге A_iA_k. При этом выполнены следующие условия:

1) $\varphi_{ik} \leq c_{ik}$, то есть поток по дуге не может быть больше ее пропускной способности;

2) сумма потоков по всем дугам, выходящим из источника равна v (равна входящим в сток)

$$\sum_{i\in\{A_0\}_-}\varphi_{0i} = \nu = \sum_{i\in\{A_n\}_+}\varphi_{in}$$

3) в стационарной сети продукты не накапливаются, поэтому для каждого узла k сумма входящих потоков равна сумме выходящих потоков:

$$\sum_{i\in\{A_k\}_+}\varphi_{ik} = \nu = \sum_{i\in\{A_k\}_-}\varphi_{kl}$$

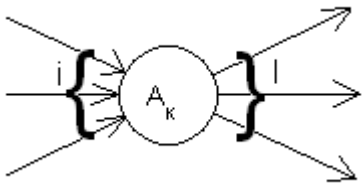


Рис. 4.5. Потоки в узле сети

Условие 3) называется *законом Кирхгофа*. Оно означает, что из каждого промежуточного узла выходит ровно столько продуктов, сколько в него поступило, то есть представляет собой математическую запись закона сохранения. Будем считать, что в сети нет равных дуг, то есть дуг, соединяющих одну и ту же упорядоченную пару узлов. В противном случае мы такие дуги объединим в одну с общей пропускной способностью.

Рассмотрим произвольный путь, ведущий от источника A_0 в сток A_n.

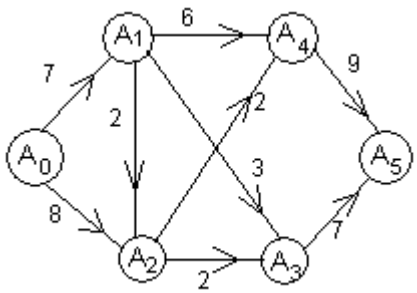


Рис. 4.6. Сеть с указанием пропускных способностей

Если ориентация всех дуг пути совпадает с направлением перемещения от узла A_0 к узлу A_n, то такой путь будем называть *ориентированным путем*. Если на пути есть противоположно ориентированные дуги, то такие дуги мы назовем обратными, в отличии от прямых дуг, направленных по пути.

Пример 3. Построим полный поток в сети, показанной на рис. 4.6. Рассмотрим путь: $A_0A_1A_4A_5$ – этот путь является ориентированным. Путь $A_0A_2A_1A_3A_5$ - не ориентирован.

Дуга сети называется *насыщенной*, если поток через нее равен пропускной способности. Поток в сети будем называть *полным*, если любой ориентированный путь из источника в сток содержит, по меньшей мере, одну насыщенную дугу. Если путь содержит не насыщенные дуги, то поток по этому пути можно увеличить до пропускной способности (минимальной) дуги в этом пути.

Начальный поток примем равным нулю. Выберем ориентированный путь, ведущий из источника A_0 в сток A_5. В нем ни одна из дуг не является насыщенной. Мы можем увеличить поток до минимального значения φ_{ik} =min(c_{ik}). Если поток при этом не будет полным, то выберем в сети следующий путь, содержащий ненасыщенные дуги и вновь увеличим поток до минимальной пропускной способности дуги в этом ориентированном пути. Поскольку сеть конечная, а пропускная способность дуг также конечна, то через конечной число ходов алгоритм завершится, и построенный поток в сети окажется полным.

1. В рассматриваемом примере это достигается следующим образом. Выберем, например, путь $A_0A_1A_4A_5$, min пропускная способность которого равна 6 (рис. 4.7).

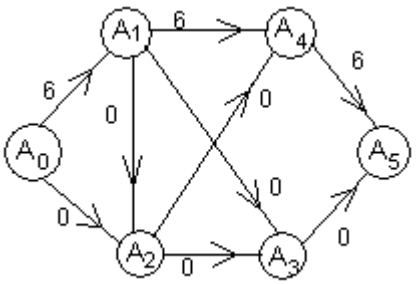


Рис. 4.7. ПОСТРОЕННЫЙ ПОТОК В СЕТИ: ШАГ 1

Новая исходная сеть после вычета потока в выбранном пути для второго шага показана на рис. 4.8.

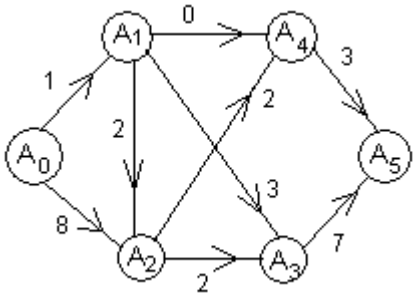


Рис. 4.8. НОВАЯ СЕТЬ: 2

2. Выберем путь $A_0A_1A_2A_4A_5$, min пропускная способность которого равна 1. Построенные потоки показаны на рис. 4.9.

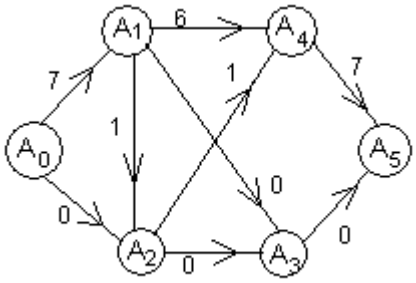


Рис. 4.9. ПОСТРОЕННЫЙ ПОТОК В СЕТИ: 3

Новая исходная сеть после вычета потока в выбранном пути для третьего шага показана на рис. 4.10.

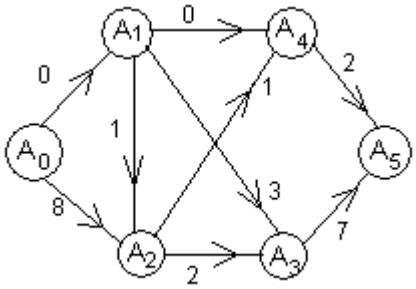


Рис. 4.10. НОВАЯ СЕТЬ: 4

3. Выберем путь $A_0A_2A_3A_5$, min пропускная способность которого равна 2. Построенный на этом шаге поток в сети показан на рис. 4.11.

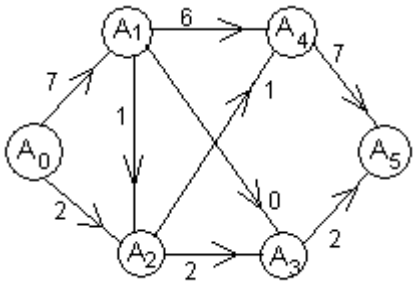


Рис. 4.11. ПОСТРОЕННЫЙ ПОТОК В СЕТИ: 5

Новая исходная сеть после вычета потока в выбранном пути для четвертого шага показана на рис. 4.12.

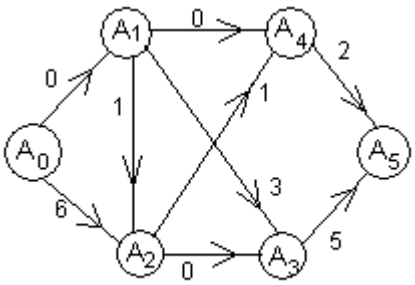


Рис. 4.12. НОВАЯ СЕТЬ: 6

4. Выберем путь $A_0A_2A_4A_5$, min пропускная способность которого равна 1. Построенный на этом шаге поток в сети показан на рис. 4.13.

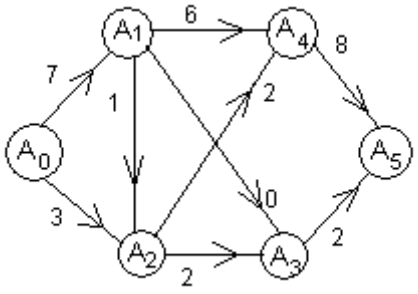


Рис. 4.13. Построенный поток: 7

Новая исходная сеть после вычета потока в выбранном пути для пятого шага показана на рис. 4.14.

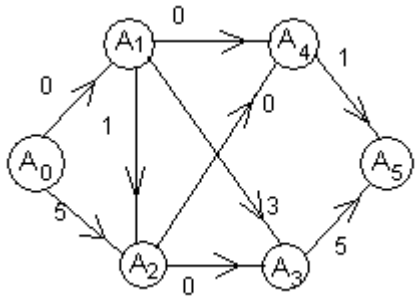


Рис. 4.14. Новая сеть: 8

Таким образом, в сети на рис. 4.14 нет путей с ненулевой пропускной способностью, и построенный поток полный: $\nu=10$.

Остается невыясненным вопрос о том, является ли этот поток максимальным из всех возможных.

Пример 4. Рассмотрим другой простой пример, для сети, показанной на рис. 4.15, и решим его двумя способами.

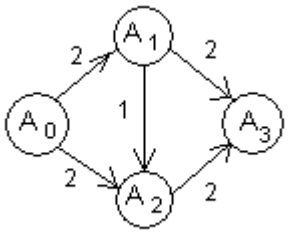


Рис. 4.15. Простая сеть

1 способ: путь $A_0A_1A_2A_3$ имеет min пропускную способность, равную 1.

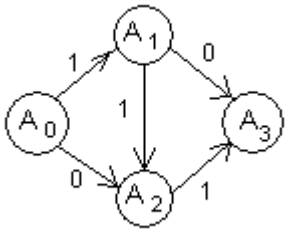


Рис. 4.16. Построенный на первом шаге поток

Для следующего шага исходная сеть имеет вид, показанный на рис. 4.17.

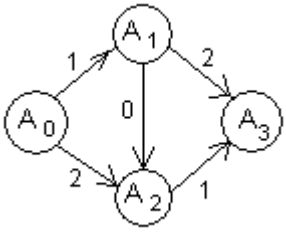


Рис. 4.17. Новая сеть

Выберем путь $A_0A_2A_3$, min пропускная способность которого равна 1.

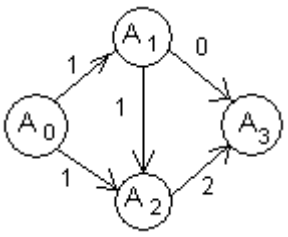


Рис. 4.18. ПОСТРОЕННЫЙ НА ВТОРОМ ШАГЕ ПОТОК

Полный поток в этом случае v = 2 (рис. 4.18).

2 способ: путь $A_0A_1A_2$ имеет min пропускную способность = 2. Путь $A_0A_2A_3$ имеет min пропускную способность = 2. Полный поток в этом случае v=4.

Как проверить, является ли построенный полный поток максимальным? Для этого имеются критерии, в основе которых лежит понятие *разреза* или *разделяющего сечения* сети. Обозначим через X некоторое множество узлов сети, соединенных с источником A_0, а $\overline{X}$ – остальные узлы сети, которые соединены с A_n. Совокупность весов всех дуг $\{X, \overline{X}\}$, начало которых принадлежит X, а конец $\overline{X}$,называется пропускной способностью разреза $c(X, \overline{X})$.

Заметим, что каждый путь из источника A_0 к стоку A_n содержит по меньшей мере одну дугу каждого разреза. Если из сети исключить все дуги какого-нибудь разреза, то не останется ни одного пути, связывающего источник A_0 со стоком A_n. Поэтому разрез сети нередко называют *разделяющим сечением*. Отсюда следует, что величина ν произвольного потока φ не превышает пропускной способности $c(X, \overline{X})$ произвольного разреза. Более того, можно доказать теорему: для любой транспортной сети величина максимального потока из источника в сток совпадает с минимальной пропускной способностью разреза. Если найденный полный поток не совпадает с максимально возможным, то его можно увеличить для прямых дуг оставшегося непрямого пути с одновременным уменьшением на ту же величину потока для обратной дуги.

4.5. Реализация сетей в трехмерном пространстве. До сих пор мы ограничивались математическими графами, и проблемы конкретной их реализации не рассматривались. Однако любая реальная физическая сеть состоят из элементов и соединяющих дуг, имеющих определенные размеры. Это накладывает определенные ограничения на возможности размещения сети в трехмерном пространстве. Соответствующая математическая задача была решена в работе [Колмогоров А.Н., Бардзинь Я.М. О реализации сетей в трехмерном пространстве // Проблемы кибернетики. 1967. Вып.19. С. 261-268.], где минимальное расстояние между дугами сети, а также между узлами ограничено. Ответ на поставленный вопрос дает следующая теорема: любая сеть с n вершинами может быть реализована в сфере радиуса $C\sqrt{n}$, где C константа, не зависящая от n. Эта теорема находится в удивительном согласии с тем фактом, что мозг устроен так, что основную его массу занимают волокна, а нейроны расположены лишь на его поверхности. Данная теорема подтверждает оптимальность в смысле объема такого строения нейронной сети. Ее, безусловно, необходимо учитывать и при попытках построения оптимальных по объему больших искусственных нейронных сетей.

4.6. Феномен «тесного мира». Социолог Гарвардского университета С.Милграм обнаружил, что каждого человека на земном шаре можно связать с другим человеком цепочкой из шести знакомых. Таким образом, несмотря на то, что на Земле живет более шести миллиардов людей, мир тесен (“small world”). Подобным свойством обладают и многие технические сети с высокой степенью кластеризации и малой средней длиной между узлами. Таким образом к выводам, следующим из теоремы Колмогорова о наиболее плотной упаковке сетей следует добавить, что если локально отказаться от требования оптимальности размещения в пространстве, а заменить это требованием быстрого установления связей между любыми узлами через малое число посредников, то наиболее эффективными становятся кластерные

структуры. Итогом может служить вывод о максимальной эффективности кластерных структур, расположенных вдоль замкнутой поверхности.

Глава 5. Разрушение сетей

5.1. Опасность разрушения сетей. Сети служат типовым представлением для широкого набора сложных систем, начиная от торговых и заканчивая водопроводными сетями, в которых наиболее важную роль играют бинарные связи. Они служат для передачи каких-либо ресурсов между узлами. Любая сеть подвержена действию ряда внутренних и внешних факторов, в результате чего ее элементы могут выйти из строя. Последствия выхода из строя отдельных элементов сети могут носить различный характер и причинять разную степень вреда системе.

Рассмотрим разрушения, которые носят характер эпидемии, когда разрушение какого-либо элемента ведет на следующем шаге к разрушению соседних с ним элементов. Пусть первоначально имеется один разрушенный узел сети, и на каждом последующем шаге происходит разрушение узлов соседних с разрушенными ранее узлами. Если начальный узел не один, то можно ввести дополнительный фиктивный узел, соединенный с узлами начального состояния, и тогда исходное состояние разрушения исходной сети возникнет после первого шага. Дальнейшую эволюцию разрушения сети легко проследит на графе визуально, если граф невелик. Для больших систем такая задача становится неразрешимой для человека, и для ее решения необходимо развить адекватный аналитический аппарат. В результате исследования желательно определить время разрушения всей сети (количество шагов до полного разрушения), время потери полной связанности сети и наметить эффективные меры защиты сети.

5.2. Алгоритм разрушения. В рассматриваемом сценарии разрушение идет по смежным узлам. Каждый шаг по времени означает добавление

одного шага к пути на графе длиной t. Это позволяет использовать для описания разрушения сети известную теорему, согласно которой, если возвести матрицу смежности графа C в степень t, то элемент $A_{ij}(t)$ матрицы $A(t) = C^t$ равен числу путей длины t из узла V_i в узел V_j. Нулевые элементы A_{ij} говорят о том, что узел V_j нельзя достичь за t шагов из узла V_i. Полное разрушение сети по прошествии времени t при старте из узла V_i означает, что в соответствующей строке суммы матриц $A(t)$ все элементы стали ненулевыми. Напомним, что для неориентированных графов исходная матрица C является действительной и симметричной, и такими же действительными и симметричными матрицами будут и все ее степени.

Для неориентированных графов в силу действительности и симметричности матрицы $\tilde{N}$ ее можно привести к виду $C = BLB^{-1}$, где L - диагональная матрица, составленная из собственных значений матрицы C:

$$L = \begin{pmatrix} \lambda_1 & 0 & \text{K} & 0 \\ 0 & \lambda_2 & & \\ \text{K} & & \text{K} & \\ 0 & \text{K} & & \lambda_n \end{pmatrix}. \qquad (5.1)$$

Тогда

$$A(t) = C^t = \left(BLB^{-1}\right)^t = BL^tB^{-1}. \qquad (5.2)$$

На больших временах доминирует элемент $(\lambda_{\max})^t$.

Если построить матрицу

$$T(t) = I + C + C^2 + \text{K} + C^t, \qquad (5.3)$$

то ее ненулевые элементы фиксируют все узлы сети, разрушенные за t шагов. Выбирая строку, которая первой станет целиком ненулевой, мы можем определить минимальное время разрушения сети $t_{\min}$ и начальную вершину V_i, которая соответствует такому варианту. Определив последнюю

строку, ставшую ненулевой, мы, тем самым, определим максимальное время разрушения сети $t_{\max}$, и найдем соответствующий стартовый узел. Таким образом, матрица $T(t)$ описывает все разрушения, происшедшие в сети за время t. Для плоских сетей можно построить изоморфный двойственный граф, заменив вершины дугами, а дуги вершинами и рассмотреть задачу последовательного разрушения связей изложенным выше методом.

Для защиты сетей могут быть разработаны различные стратегии, в зависимости от преследуемой цели. В качестве таковой цели может быть максимальное замедление разрушения, путем увеличения длины путей, то есть времени t. С точки зрения структуры матриц $T(t)$ защита означает вычеркивание наиболее быстро заполняемых строк, то есть защиту от разрушения соответствующих элементов, что будет удлинять возможные пути и увеличивать время полного разрушения.

5.3. Защита сети. Для прояснения возможных вариантов защиты сети полезно рассмотреть предельные частные случаи. Так наиболее трудно, практически невозможно, защитить полносвязную сеть, так, как любой узел достижим за один шаг. В радиально-кольцевом графе целесообразнее всего защитить центральный узел, что увеличит время разрушения сети в $N/2$ раз, где N - число узлов на кольце. Следует иметь в виду, что если защита при этом выводит узел из функционирования сети, то характерное время последовательного доступа между элементами сети увеличивается от 2 через центр до $N/2$ при доступе по кольцу. Если сеть представляет собой дерево, то пути разрушения не сходятся вместе, и полное разрушение сети идет достаточно долго при любой начальной точке разрушения.

Обеспечение качественной работы сети предполагает выполнение противоречивых условий. С одной стороны, путь между любыми двумя узлами должен быть максимально коротким. Это обеспечивает быстроту перемещения ресурсов по сети. С другой стороны, эта быстрота связей одновременно ведет к возможности быстрого разрушения сети.

Противоречие можно устранить, если иметь в виду две сети: одну сеть – для нормального функционирования, а вторую, – функционирующую в режиме защиты и являющуюся подграфом первой. Тогда исходная сильносвязная сеть при угрозе разрушения трансформируется в дерево. Поскольку порождающих деревьев может быть много, то из них имеет смысл выбрать самое устойчивое. Дерево характерно тем, что в нем разрушение носит характер ветвящегося процесса, а полное время разрушения соответствует самой длинной ветви, идущей от начального узла. У каждого дерева есть центр, то есть узел, который максимально удален от концов. Для его нахождения выбирается самый длинный путь из данного узла i до всех возможных висячих вершин j. Затем из всех вершин i выбирается та, для которой эта величина минимальна:

Далее из всех графов нужно выбрать тот, для которого полученная величина максимальна:

$$\tilde{t} = \max_k \min_i \max_j \left(t_{ij}^k\right). \qquad (5.4)$$

Индекс k нумерует разные порождающие графы. Центр полученного дерева нужно защитить.

Для анализа очень больших сетей можно перейти к непрерывному пределу. В этом случае матрица смежности C_{ij} заменяется симметричным ядром $C(x, y)$, а последовательные шаги разрушения описываются итерациями

$$\begin{aligned} \tilde{\tilde{N}}(x, y, t) &= C(x, y) + \int C(x, y_1) C(y_1, y) dy_1 + \text{K} \\ &+ \int C(x, y_1) C(y_1, y_2) \text{K}\ C(y_{t-1}, y) dy_1 \text{K}\ dy_{t-1}. \end{aligned} \qquad (5.5)$$

Описанный метод обобщается на описание разрушения ориентированных сетей. При этом матрицы перестают быть симметричными.

Эффективность работы и надежность сети являются взаимно дополнительными и не могут быть обеспечены одновременно. Имеются

простые и эффективные матричные методы моделирования и анализа процесса разрушения сетей. Это позволяет под новым углом взглянуть на глобализацию как проблему эффективности и надежности глобальных сетей информации, финансов и распространения эпидемий.

Глава 6. Принятие решений при неопределенности целей

6.1. Противоречивость целей. Назначение цели в задачах исследования операций и формализация цели, то есть выбор целевой функции почти всегда является трудной проблемой. Часто бывает необходимо выбрать стратегию, которая делает максимальным доход при минимальных затратах. Таким образом мы имеем одновременно две задачи

$$\begin{aligned} & f(x) \to \max \\ & -F(x) \to \max \ \ (F(x) \to \min) \end{aligned} \qquad (6.1)$$

где $f(x)$ *и* $F(x)$ – функции, характеризующие соответственно доход и затраты.

Такая задача, как правило, решений не имеет. В принципе для увеличения дохода необходимо увеличивать затраты. Таким образом, поставленные цели противоречат друг другу. Для того, что бы свести данную задачу исследования операций к стандартной задаче оптимизации, необходимо сформулировать дополнительные гипотезы, не вытекающие прямо из постановки задачи.

Остановимся на некоторых наиболее употребительных способах преодоления неопределенности целей, когда стоит задача обеспечить максимальное значение функциям $f_1(x)$, $f_2(x)$, …, $f_n(x)$ одновременно.

6.2. Линейная свертка. Вместо n отдельных критериев часто рассматривают один критерий вида

$$F(x) = \sum_{i=1}^{n} c_i f_i(x) \ \ , \qquad (6.2)$$

где c_i – некоторые положительные числа, обычно нормированные условием $\sum_{i=1}^{n} c_i = 1$.

Коэффициенты c_i являются результатами экспертизы. Они отражают представление оперирующей стороны о содержании компромисса, который она вынуждена принять. Таким образом происходит ранжирование целей на основе назначения весовых коэффициентов.

6.3. Использование контрольных показателей. Часто в задачах планирования и проектирования задается некоторая система нормативов: F_1, F_2, ...,F_n . Это значит, например, что функции $f_i(x)$ должны достигать максимальных значений при условиях

$$f_i(x) \geq F_i, \ \ i = 1, 2, \text{K}\ n \tag{6.3}$$

В таких случаях целевую функцию можно задать в виде

$$F(x) = \min \frac{f_i(x)}{F_i} \tag{6.4}$$

и искать вектор x, который обеспечивает максимальное значение $F(x)$. Смысл процедуры прост. Критерий (6.4) дает значение наихудшего показателя, а условие $F(x) \to \max$ делает его максимально хорошим.

Если на компоненты вектора x положены линейные ограничения вида

$$\sum_i a_{ij} x_j \leq b_i \tag{6.5}$$

а функции $f_i(x)$ также линейные функции

$$f_i(x) = \sum_j d_{ij} x_j \ , \tag{6.6}$$

то задача сводится к линейному программированию.

6.4. Простейший способ преодоления неопределенности целей. Если система контрольных показателей имеет вид (3), то в ряде случаев выбирают

основной критерий, например $f_1(x)$. Тогда снова получается однокритериальная задача

$$f(x) \to \max \qquad (6.7)$$

при условиях (6.3).

9.5. Метрика в пространстве целевых функций. Предположим, что мы решили систему однокритериальных задач

$$f_i(x) \to \max, \ \ i = 1, 2, \text{K}\ n$$

и нашли в i-ой задаче вектор $x=x_i$, доставляющий максимальное значение критерию $f_i(x)$. Совокупность скалярных величин $f_i(x_i)$ определяют в пространстве критериев некоторую точку, которую назовем абсолютным максимумом.

Если векторы x_i различны, то такой абсолютный максимум недостижим. Однако, можно поставить задачу о максимальном приближении к нему. В качестве меры близости введем расстояние в пространстве критериев, например, в виде евклидова расстояния

$$h = \sqrt{\sum_i (f_i(x) - f_i(x_i))^2} \ . \qquad (6.8)$$

Тогда h есть расстояние от точки $(f_1(x), f_2(x), \ldots, f_n(x))$ до точки $(f_1(x_1), f_2(x_2), \ldots, f_n(x_n))$ в пространстве критериев.

В качестве нового скалярного критерия можно теперь принять функцию (6.8). Ее минимизация позволяет ближе всего подойти к абсолютному максимуму.

6.6. Компромиссы Парето. К анализу многокритериальных задач можно подойти и с несколько иных позиций, попытавшись исключить заведомо плохие решения. Один из путей решения подобных задач предложен итальянским экономистом В. Парето в 1904г.

Пусть x_0 соответствует некоторому выбору, и имеется лучший выбор x. Тогда

$$f_i(x) \geq f_i(x_0),\ i = 1,2,\text{K},n. \qquad (6.9)$$

Ясно, что выбор x предпочтительнее x_0. Множество всех таких x называют *множеством Парето*. Вектор x называют *наилучшим вектором результатов* или *вектором Парето*, если из $f_i(x) \geq f_i(x_0)$ для любого i следует $f_i(x) = f_i(x_0)$.

Предположим, что цели определяются двумя однозначными функциями

$$\begin{matrix} f_1(x) \to \max \\ f_2(x) \to \max \end{matrix}. \qquad (6.10)$$

Тогда каждому допустимому значению x отвечает одна точка на плоскости (f_1, f_2) (рис. 6.1).

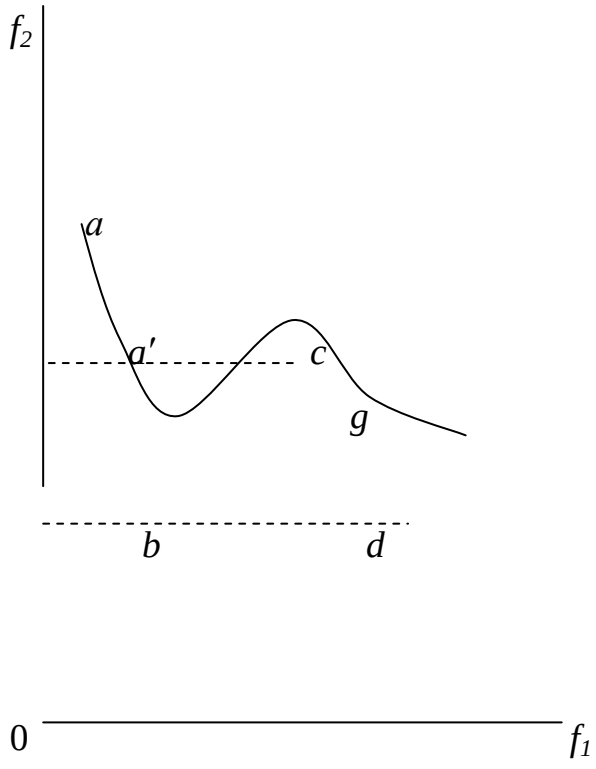


Рис. 6.1. МНОЖЕСТВО ПАРЕТО ДЛЯ СЛУЧАЯ ДВУХ КРИТЕРИЕВ

Равенства $f_1 = f(x)$, $f_2 = f(x)$ параметрически задают некоторую кривую *abcd* в этой плоскости. Участок *bc* не принадлежит множеству Парето,

поскольку на этом участке одновременно растут обе величины f_1 и f_2. На этом основании исключается и участок $a'b$, поскольку для любой точки e из участка $a'b$ найдется точка g участка cd, в которой обе функции больше, чем в точке e. Значит, множеству Парето могут принадлежать только участки aa' и cd.

В теории принятия решений существует *принцип Парето*, согласно которому выбирать в качестве решения можно только векторы x из множества Парето. Принцип Парето не определяет решение, но сужает множество возможных решений. Построение множеств Парето облегчает поиск компромисса в многокритериальных задачах.

Глава 7. Динамическое программирование

7.1. Принцип оптимальности. В целом ряде случаев операции разбиваются на ряд этапов. На каждом этапе имеется управление операцией, от результатов которой зависит, как выполнение данной операции, так и операции в целом. *Динамическое программирование* представляет собой метод оптимизации процесса поэтапного управления.

Оптимизация многошагового процесса управления проводится на основе *принципа оптимальности*. Он состоит в том, что планирование проводится так, чтобы учитывался результат не только на данном этапе, но и общий результат в конце многошаговой операции.

Управление u всей операцией представляет собой совокупность шаговых управлений $u=(u_1,u_2, \ldots u_m)$, где m – число шагов. Эффективность операции (доход) F зависит от выбранной системы управлений, то есть является функцией от u:

$$F=F(u). \qquad (7.1)$$

Задача состоит в таком выборе управления, чтобы F была максимальной. Управление u^*, при котором $F=F_{max}$ называется оптимальным. Идея решения такой задачи состоит в постепенной оптимизации. На каждом шаге оптимизация должна проводиться с учетом влияния всех последующих

шагов на конечный результат. Исключение составляет последний шаг. Он может быть спланирован оптимально независимо от предыдущих шагов. Затем можно оптимизировать предпоследний шаг и т.д.

Процесс *динамического программирования* выстраивается от конца к началу. На каждом шаге ищется оптимальное продолжение процесса относительно достигнутого в данный момент состояния системы, и такое управление называется условно оптимальным. При оптимизации управления многошаговый процесс проходится дважды:

1. от конца к началу, в результате чего находят условно-оптимальные управления на каждом шаге и условный оптимальный выигрыш на всех шагах;
2. от начала к концу, в результате чего определяют оптимальные шаговые управления.

7.2. Задача о распределении ресурсов. Пусть имеется начальное количество средств k_0, которое нужно распределять в течение m лет между производствами I и II. Если средства X вложить в первое производство, то за год будет получен доход $f(X)$, средства уменьшатся, и к концу года от них останется некоторая часть $\varphi(X)<X$. Аналогично, средства Y, вложенные в отрасль II, принесут за год прибыль $g(Y)$ и уменьшатся до величины $\psi(Y)<Y$. На следующий год вновь осуществляется вложение уже оставшихся средств без поступлений извне и без использования прибыли. Требуется найти способ управления ресурсами, при котором суммарная прибыль обеих производств за m лет будет максимальной.

Задача может быть решена методом динамического программирования. Естественным шагом процесса является в данном случае год. Управление u_i на i-ом шаге состоит в выделении средств X_i и Y_i, вкладываемых в производства I и II. Управление операцией состоит из совокупности шаговых управлений за m лет.

Состояние системы перед i-м шагом характеризуется одним параметром k_{i-1} – количеством оставшихся к этому времени средств. Управление u_i на i-ом шаге заключается в том, что первому производству выделяются средства X_i. Тогда второму производству остается величина $Y_i=k_{i-1}-X_i$. Прибыль на i-ом шаге $W_i=f(X_i)+g(k_{i-1}-X_i)$.

Обозначим через $W_j(k)$ максимальную прибыль, которую можно получить, начиная с j-того года по последний год включительно, при условии, что к началу j-того периода осталось k средств. Если в начале i-того периода выделено X_i средств в первое производство, то максимальная прибыль, которая может быть получена, начиная с i-того года, будет равна сумме прибыли за i-тый год и максимальной прибыли за оставшиеся годы, т.е.

$$W_i(k)=f(X_i)+g(k_{i-1}-X_i)+W_{i+1}(\varphi(X_i)+\psi(k_{i-1}-X_i)). \qquad (7.2)$$

Максимальная прибыль за годы, начиная с i-того, будет получена, если выбрать X_i так, чтобы функция (7.2) принимала максимальное значение. Тем самым получаем связь между максимальной прибылью, после с (i-1) –го года $W_i(k)$ и, максимальной прибылью, которую можно получить после i-того года в виде функционального уравнения

$$W_i(k)=\mathbf{max}_{0\le X_i\le k}[f(X_i)+g(k-X_i)+W_{i+1}(\varphi(X_i)+\psi(k-X_i))]. \qquad (7.3)$$

Это так называемое *основное функциональное уравнение динамического программирования*. Поскольку величина k_i заранее неизвестна, в уравнении она заменена на k. Значение $W_i(k)$ нужно находить при всех k. Далее нужно найти значение $X_i(k)$ и соответствующее ему значение $Y_i(k)=k-X_i(k)$, при

котором достигается оптимальное уравнение на i-том шаге (при условии $k_i = k$). Величина $W_i(k)$ является условно-оптимальной прибылью.

Условно-оптимальная прибыль на последнем шаге

$$W_m(k) = \max_{0 \le X_m \le k} \{f(X_m) + g(k - X_m)\}. \qquad (7.4)$$

Ей соответствует условно-оптимальное управление $X_m(k)$, на котором достигается максимум. Для построения оптимального управления на предыдущих этапах воспользуемся уравнением (3), где W_{i+1} теперь имеет условно определенное значение. В итоге получаем условно оптимальные управления $X_{m-1}(k)$, $X_{m-2}(k)$, $\ldots X_1(k)$. Чаще всего функции $W_i(k)$ и $X_i(k)$ находятся в табличном виде.

Теперь можно по известному начальному запасу средств k_0 определить максимальную прибыль за весь период: $W_{max} = W_1(k_0)$. Затем можно найти оптимальное управление на первом шаге $X_1^* = X_1(k)$ и $Y_1^* = k_0 - X_1^*(k_0)$. По окончании первого этапа останутся средства в размере $k_1^* = \varphi(X_1^*) + \psi(Y_1^*)$. Находим затем оптимальное управление на втором шаге: $X_2^* = X_2(k_1^*)$, $Y_2^* = k_1^* - X_2(k_1^*)$ и остаток средств $k_2^* = \varphi(X_2^*) + \psi(Y_2^*)$ к концу второго года. Продолжая этот процесс, получим последовательность значений

$$k_0 \to X_1(k_0) \to k_1^* \to X_2(k_1^*) \to k_2^* \to \ldots$$
$$\to k_{m-1} \to x_m(k_{m-1}^*) \to k_m^*.$$

Величина k_m^* дает количество средств, оставшихся к концу планируемого периода.

В первое производство вкладывается по годам $X_1^*, X_2^*, \ldots, X_m^*$ средств, во второе производство $Y_1^* = k_0 - X_1^*$, $Y_2^* = k_1 - X_2^*$,…,
$Y_m^* = k_{m-1}^* - X_m^*$ средств.

Глава 8. Элементы теории игр

8.1. Конфликты как игры. В природе и обществе часто возникают конфликтные ситуации, в которых участвуют стороны с различными и даже с противоположными интересами. Конфликтные ситуации возникают в самых разных областях деятельности: при операциях купли-продажи, в большом и малом бизнесе, в спорте и т.д.

Математическая теория конфликтных ситуаций называется *теорией игр*. Ее задачей является выработка рекомендаций поведения, которое приводило бы к наибольшей выгоде той или иной стороны. Игры различаются по числу сторон на *парные* и *множественные*.

Мы рассмотрим только парные игры. Стороны назовем игроками A и B. Для описания игры необходимо сформулировать ее правила, объем информации каждой стороны о поведении другой и результат игры, к которому приводит каждая последовательность ходов. Парная игра называется *игрой с нулевой суммой*, если выигрыш одного игрока равен проигрышу другого. Мы будем рассматривать только такие игры.

Стратегией игрока называется совокупность правил, определяющих выбор его действий при каждом ходе в зависимости от сложившейся ситуации. Задача теории игр заключается в выборе стратегии, приводящей к наибольшему выигрышу в предположении, что второй игрок также придерживается наилучшей стратегии.

Пусть игрок А имеет возможность применять m стратегией A_1, A_2 ... , A_m, а игрок B – n стратегий $B_1, B_2, \ldots, B_n$. Обозначим через a_{ij} выигрыш игрока A, если он выберет стратегию A_i, а игрок B – стратегию B_j. В этом случае игра полностью описывается матрицей a_{ij}, которая называется *платежной матрицей* или *матрицей игры*. Выигрыш игрока A означает проигрыш игрока B, для которого платежная матрица $b_{ij} = -a_{ij}$. Предположим, что игрок A выбрал стратегию A_i. Естественно, что игрок B выберет стратегию, при

которой выигрыш α_i игрока A будет минимальным, то есть $\alpha_i = \min\limits_i a_{ij}$. В свою очередь игрок A предпочтет стратегию A_k, при которой выигрыш будет наибольшим и $\alpha_k = \max\limits_j a_{ij} = \max\limits_i \min\limits_j a_{ij}$. Стратегия A_k игрока A называется *максиминной.* Она гарантирует игроку A выигрыш α_k. Как бы не играл игрок B выигрыш игрока A не будет меньше этого числа. Величина α_k называется *нижней ценой игры*.

Рассмотрим теперь поведение игрока B. Если он выберет стратегию B_j, то A примет стратегию, приводящую к наибольшему выигрышу $\beta_j = \max\limits_i a_{ij}$. Тогда для B следует выбрать такую стратегию, при которой выигрыш A будет наименьшим, то есть $\beta_l = \min\limits_j \max\limits_i a_{ij}$ Стратегию B_l называют *минимаксной* , и она гарантирует игроку B, что его проигрыш будет не более величины β_l как бы ни играл игрок A. Величина β_l называется *верхней ценой игры*.

8.2. Основное неравенство и игра с седловой точкой. По определению $\alpha_k = \min\limits_j a_{kj}$, то есть α_k не больше любого элемента из k-ой строки. Поэтому $\alpha_k \le \alpha_{kl} \le \max a_{il} = \beta_l$. Отсюда следует $\alpha_k \le \beta_l$ – основное неравенство. Из основного неравенства вытекает ряд важных следствий.

1. Нижняя цена игры не превосходит верхней цены игры.

2. Если игрок A придерживается максиминной стратегии, а игрок B минимаксной стратегии, то выигрыш a_{kl} заключен между нижней и верхней ценами игры $\alpha_k \le a_{kl} \le \beta_l$.

3. Если нижняя и верхняя цены игры совпадают ($\alpha_k = \beta_l$), то основное неравенство превращается в равенство и элемент α_{kl} является минимальным в строке и максимальным в столбце. При этом по аналогии с непрерывными функциями говорят, что игра имеет *седловую точку*. Справедливо и *обратное утверждение*: если игра имеет седловую точку, то нижняя и верхняя цены совпадают.

Если игра имеет седловую точку и игрок A придерживается максиминной стратегии, то игроку B невыгодно отклоняться от минимаксной стратегии и обратно, если B придерживается минимаксной стратегии, то A невыгодно отклоняться от максиминной стратегии. Однако имеются игры без седловых точек, для которых сформулированные утверждения неверны.

8.3. Игры с вероятностным выбором стратегии. Часто игра может происходить многократно и игроки в процессе игры могут менять стратегии. Предположим, что игрок A применяет стратегии A_i с заданными вероятностями p_i. Поскольку стратегии A_i представляют собой полную систему событий, то

$$\sum_{i=1}^{m} p_i = 1. \qquad (8.1)$$

В свою очередь игрок B применяет стратегии B_j с частотами (вероятностями) q_j и

$$\sum_{j=1}^{n} q_j = 1. \qquad (8.2)$$

Наборы чисел $(p_1,\ldots,p_m)$ и $(q_1,\ldots,q_n)$ называются *смешанными стратегиями*. Если некоторое $p_i=1$, а все остальные вероятности нулевые, то это означает, что игрок A постоянно придерживается одной стратегии A_i. Такая стратегия называется *чистой*. Аналогично определяется чистая стратегия B игрока B_j, когда $q_j=1$.

Решением игры в смешанных стратегиях называется такая пара стратегий игроков A и B, называемых оптимальными, что если один из них придерживается своей оптимальной стратегии, то другому также невыгодно отклоняться от этой стратегии.

Выигрышем игрока A естественно назвать среднее значение выигрыша на множестве реализуемых стратегий. Если игроки A и B играют независимо, то вероятность того, что игрок A примет стратегию A_i, а игрок B – стратегию B_j есть произведение вероятностей этих независимых событий p_i, q_j.

Поэтому среднее значение v выигрыша игрока A равно математическому ожиданию

$$v = \sum_{i,j} a_{ij} p_i q_j \,. \qquad (8.3)$$

Если используются оптимальные смешанные стратегии, то соответствующее значение v называется *ценой игры*.

Существует связь между теорией игр и линейным программированием, методы которого можно использовать для анализа игр в смешанных стратегиях. Пусть все числа a_{ij} положительные. Если это не так, то к элементам матрицы a_{ij} можно прибавить такое положительное число c, что все элементы матрицы станут положительными. Зафиксируем некоторую смешанную стратегию $(p_1, p_2,\ldots,p_m)$ игрока A. Число ζ назовем *гарантированным выигрышем игрока A* при выбранной стратегии, если выигрыш A не будет меньше ζ при любой стратегии игрока B. Пусть B выбирает чистую стратегию (1,0,…,0). Тогда по формуле выигрыш

$$v = a_{11} p_1 + a_{21} p_2 + \ldots + a_{m1} p_m \,.$$

По свойствам величины ζ

$$a_{11} p_1 + a_{21} p_2 + \ldots + a_{m1} p_m \geq \zeta \,.$$

Проверив все чистые стратегии игрока B, приходим к системе неравенств

$$a_{1j} p_1 + a_{2j} p_2 + \ldots + a_m p_m \geq \zeta \,. \qquad (8.4)$$

Если неравенства (8.4) выполнены, то при любой смешанной стратегии $(q_1, q_2,\ldots,q_n)$ игрока B выигрыш

$$v = \sum_{i,j} a_{ij} p_i q_j = \sum_j \left(\sum_i a_{ij} p_j \right) q_j \geq \sum_j \zeta q_j = \zeta \sum_j q_j = \zeta, \qquad (8.5)$$

То есть $\nu > \zeta$ и величина ζ является гарантированным выигрышем. Выполнение неравенств (4) является необходимым и достаточным условием того, чтобы величина ζ была гарантированным выигрышем.

Естественно в качестве оптимальной стратегии A принять ту, которая дает максимальный гарантированный выигрыш ζ. Введем обозначения $x_i = p_i / \zeta$, тогда неравенства (8.4) примут вид

$$\sum a_{ij} x_j \geq 1. \qquad (8.6)$$

Равенство (8.1) переходит в

$$\sum_i x_i = 1/\zeta = F. \qquad (8.7)$$

Таким образом, задача заключается в минимизации линейной формы (8.7) при ограничениях (8.6). Тем самым она приняла стандартный вид для линейного программирования. Подобным же образом может быть поставлена и решена задача о наименьшем неизбежном проигрыше η игрока B. Для оптимальных стратегий $\zeta = \eta$. Тем самым мы приходим к основной *теореме Неймана теории игр*: всякая конечная игра имеет решение.

8.4. Выбор стратегии. Бывают конфликтные ситуации, в которых одна из сторон действует неопределенно. Игру такого типа называют игрой с природой. Пусть игрок B теперь является природой. Если игрок А выбирает чистую стратегию A_0, то математическое ожидание выигрыша равно $\sum a_{ij} q_j$. Поэтому наиболее выгодной будет та стратегия, при которой достигается

$$\max \sum_j a_{ij} q_j . \qquad (8.8)$$

Если информация о состояниях природы мала, то разумно считать, что все состояния природы равновероятны. Тогда игрок А выбирает ту стратегию, при которой достигается

$$\max \frac{1}{n} \sum_{j=1}^{n} a_{ij} . \qquad (8.9)$$

Критерий (8.9) выбирает стратегию, для которой среднее арифметическое элементов соответствующей строки максимальное.

Существуют и другие виды критериев для выбора наилучшей стратегии. По *критерию Вальда* выбирается стратегия, обеспечивающая $\max\limits_i \left(\min\limits_j a_{ij} \right)$. Этот критерий предполагает, что природа будет действовать наихудшим для A способом.

Критерий Гурвица позволяет смягчить требования критерия Вальда, и записать их в виде

$$\max_i \left[\lambda \min_j a_{ij} + (1-x) \max_j a_{ij} \right], \qquad (8.10)$$

где $0 \le \lambda \le 1$.

Критерий Сэвиджа основан на минимальности *рисков* $r_{ij} = \max\limits_i a_{ij} - a_{ij}$ (разницей между наибольшим выигрышем игрока A и самим выигрышем. Критерий запишем в виде

$$r = \min(\max_i a_{ij} - a_{ij}). \qquad (8.11)$$

Глава 9. Генетические алгоритмы и эволюционное программирование

9.1. Генетические понятия. Генетические алгоритмы используется для оптимизации сложных систем. Идея генетических алгоритмов основана на эволюционной теории, согласно которой эволюция живых организмов определяется следующими факторами:

1. случайная изменчивость;
2. наследственность;
3. естественный отбор.

Определим ряд основных понятий, используемых в теории генетических алгоритмов.

Ген - единица наследственной информации, не делимая в функциональном отношении, которая передается от родителей к потомкам.

Аллель - фиксированная форма гена.

Геном – совокупность всех генов данного организма.

Мутация – случайное наследственное изменение отдельного гена.

Естественный отбор – процесс, направленный на повышение вероятности оставления потомства одной формы организма, по сравнению с другими.

Изменчивость – разнообразие признаков и свойств у особей групп особей любой степени родства.

Популяция - совокупность особей определенного вида, внутри которого осуществляется случайное скрещивание.

Селекция – форма искусственного отбора, при котором эволюция направляется факторами внешней среды.

Эволюция – процесс постепенного и непрерывного изменения форм организмов от одного состояния к другому.

9.2. Генетический алгоритм представляет собой адаптивный поисковый метод, основанный на селекции лучших элементов в популяции подобно эволюционной теории Ч. Дарвина. Основой генетических алгоритмов служит модель биологической эволюции и методы случайного поиска. Эволюционный поиск с точки зрения приобретения информации – это последовательное преобразование одного конечного множества промежуточных решений в другое. Цель разработки генетических алгоритмов состоит в том, чтобы понять механизмы развития и адаптации естественных биологических и интеллектуальных систем, а также использовать эволюционные модели для решения научно-технических задач оптимизации. Задача оптимизации понимается как поиск абсолютного max(min) некоторой целевой функции.

Все генетические алгоритмы работают на основе начальной информации, в качестве которой выступает множество исходных

альтернативных решений P, которые называются *исходной популяцией*. Популяция $P^t=\{P_1, P_2,\ldots, P_n\}$ - множество элементов P_i соответствующих некоторой генерации генетического алгоритма t =0,1,2.. а N – размер популяции. Каждый элемент популяции P_i представляет собой одну или несколько хромосом индивидуальной особи, то есть одно или несколько альтернативных решений.

Хромосомы состоят из генов $P_i=\{g_1, g_2, \ldots, g_v\}$ которые составляют части закодированного решения. Позиция определенного гена в хромосоме называется *локусом*. Функциональное назначение генов определяет аллель. Гены могут иметь числовые или функциональные значения. Генетический материал хромосом обычно кодируется на основе двоичного кода {0,1}: P_i = {001 001 101}.

Элементы популяции генетических алгоритмов часто называют *родителями*. Родители выбираются из популяции на основе заданных правил, а затем смешиваются (скрещиваются) для производства потомства. Дети и родители в результате генерации, то есть одного цикла эволюции, создают новую популяцию. *Генерация* – это процесс реализации одной итерации алгоритма, которая называется поколением.

9.3. Эволюция в популяции – это процесс чередования поколений, в ходе которого хромосомы изменяют свои значения так, что в результате каждое новое поколение наилучшим образом приспосабливается к внешней среде. Каждый элемент популяции имеет определенный уровень качества, который характеризуется значением целевой функции (функции полезности). Эта функция используется в генетических алгоритмах для сравнения между собой альтернативных решений и выбора лучших.

Основная задача генетических алгоритмов состоит в оптимизации целевой функции. Каждая популяция обладает наследственной изменчивостью, что означает возможность случайных отклонений в генах и хромосомах в каждом поколении. При этом наследственные признаки закрепляются, если они имеют приспособительный характер, то есть

обеспечивают большее значение целевой функции. Отбраковка менее приспособленных потомков составляет суть селекции. Генетический алгоритм обеспечивает также адаптацию к изменяющейся окружающей среде, то есть к изменяющейся целевой функции.

При использовании традиционных методов оптимизации всякую новую изменившуюся задачу обычно приходится решать заново. При эволюционном подходе оптимизация часто может быть продолжена с помощью использования механизмов дополнения и видоизменения популяции. Генетический алгоритм обладает также тем достоинством, что он сходится к результатам значительно быстрее, чем простые алгоритмы случайного поиска типа Монте-Карло. Напомним, что алгоритм Монте-Карло заключается в выборе примеров случайным образом и сравнение качества альтернатив между собой.

9.4. Канонический генетический алгоритм состоит в выполнении следующих шагов:

1. Задается функция $f(P_i)$, определяющая эффективность каждого, найденного решения при значениях параметров решения P_i. P_i кодируется как вектор, который называется хромосомой. В хромосоме, каждый элемент (элемент вектора) представляет собой ген. Ген кодируется в двоичном представлении.

2. В соответствии с ограничениями, налагаемыми на параметры условиями задачи, инициализируется исходная популяция P^0 потенциальных решений, состоящая из некоторого количества хромосом N, число которых задается в начале работы алгоритма и в процессе эволюции обычно не меняется.

3. Каждой хромосоме в популяции, на основе вычисления целевой функции $f(P_i)$, присваивается вероятность воспроизведения p_i. Одним из простейших способов определения p_i является пропорциональный отбор, при котором:

$$p_i = \frac{f(p_i)}{\sum_{j=1}^{N} f(p_j)}. \qquad (9.1)$$

4. В соответствии с вероятностями воспроизведения p_i создается новая популяция хромосом, причем с большей вероятностью воспроизводятся наиболее эффективные элементы. Хромосомы производят потомков, используя операцию рекомбинации. Операция *рекомбинации* состоит из двух операторов – кроссинговера, при котором хромосомы скрещиваются, обмениваясь частями строк, и *оператора мутации*, который осуществляет вероятностные изменения генов.

5. Генетический алгоритм останавливается, если получено удовлетворительное решение, т.е. найдено решение с заданной точностью или закончилось время, отведенное на эволюцию (параметр t вышел за допустимую границу).

9.5. Оператор кроссинговера производит скрещивание хромосом и обмен генетическим материалом между родителями, для получения потомков. Этот оператор служит для исследования новых областей, пространства параметров и улучшения существующих параметров, обеспечивая эволюционное приспособление.

Простейший одноточечный кроссинговер производит обмен частями, на которые хромосома разбивается точкой, выбираемой случайно (на рис. 9.1 показан пример).

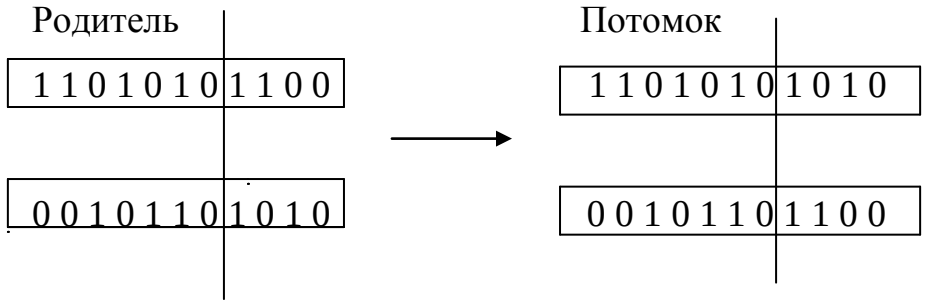


Рис. 9.1. Пример кроссинговера

Двухточечный кроссинговер обменивает куски строк, попавшие между двумя точками.

9.6. Оператор мутации применяется к каждому биту информации хромосомы с малой вероятностью $p=10^{-3}$ и меньше, в результате чего бит изменяет значение на противоположное: $0\to1,\ 1\to0$. Мутации нужны для расширения пространства поиска, то есть для предотвращения полной потери разнообразия в результате получения чистых аллелей - популяции, в которой один или несколько генов одинаковы у всех элементов популяции.

Наиболее часто в генетических алгоритмах используется бинарная кодировка для представления элемента популяции. Такой выбор связан с тем, что двоичный код наиболее прост в обработке компьютером. Существуют альтернативные двоичному коду схемы кодировки, в том числе использование представления чисел с помощью плавающей запятой при фиксированной разрядности. В зависимости от свойств конкретной решаемой задачи предпочтительными могут быть те или иные способы кодирования. Отметим также, что довольно часто применяются коды Грея.

Коды Грея построены таким образом, что каждый символ кодируется при помощи четырех бинарных символов, как показано на рис. 9.2.

0	0	0	1

0	0	1	0

1	0	0	1

Рис. 9.2. Примеры элементов кода Грея

Рассмотрим последствия работы генетического алгоритма. Пусть отбор производится с вероятностью, определяемой формулой (9.1). Если обозначить среднее значение целевой функции в популяции как

$$\bar{f} = \frac{\sum_{i=1}^{N} f_i}{N}, \qquad (9.2)$$

где N - число хромосом в популяции, то

$$p_i = \frac{1}{N}\frac{f_i}{\bar{f}}. \qquad (9.3)$$

Пусть отбираются хромосомы определенного вида, обладающего характеристикой H, и на шаге с номером t соответствующих хромосом $m(H,t)$ в результате парного обмена генетической информацией со всеми элементами популяции получится $m(H,t)\cdot N$ потомков. Отбор в соответствии с вероятностным правилом (9.3) оставит на $t+1$ шаге времени

$$m(H,t+1) = m(H,t)\cdot N \cdot \frac{f(H)}{\bar{f}} \qquad (9.4)$$

особей, обладающих свойством H. Пусть целевая функция шаблона H превышает среднее значение целевой функции по популяции. Тогда можно считать $f(H)/\bar{f} = 1+c$, где $c>1$ - константа. Тогда за t шагов, начиная с $t=0$, число хороших хромосом популяции составит

$$m(H,t) = m(H,0)\cdot(1+c)^t. \qquad (9.5)$$

С течением времени число хороших хромосом популяции растет экспоненциально, а число хромосом со значением целевой функции ниже среднего экспоненциально убывает. Это свойство известно как правило репродукции Д.Холланда. Оно является следствием обмена генами между особями популяции.

При выборе метода отбора в генетических алгоритмах возникает ряд проблем. Так, чаще всего используют пропорциональный отбор на основании формулы (9.1). Однако если добавить к любой целевой функции ограниченной вариации достаточно большую произвольную константу, то можно сделать все вероятности p_i практически одинаковыми. Это фактически уничтожает отбор, приводя эволюцию к случайному выбору. Для

устранения данного недостатка производится замена: $f(P_i) \rightarrow f(P_i) - f(P_i')$, где $f(P_i')$ – наименьшее значение целевой функции популяции.

Еще одна из математических проблем, связанная с пропорциональным отбором, состоит в том, что такая процедура не может гарантировать асимптотическую сходимость в пределе большого числа поколений к глобальному оптимуму. Наилучшая хромосома в популяции может быть потеряна в любом поколении. Поэтому результаты эволюции, достигнутые в ряде поколений, могут быть утрачены. Одним из способов преодоления этого явления является использование элитного отбора, который всегда сохраняет наилучшую хромосому в популяции.

Часть 4. Методы искусственного интеллекта

Глава 1. Искусственный интеллект: истоки и содержание проблемы

1.1. Общие представления. Под искусственным интеллектом понимают стратегии и методы решения сложных проблем. Искусственный интеллект можно определить, как область компьютерной науки, занимающейся автоматизацией разумного поведения. *Разумное поведение* не имеет единого определения, одно из них – стандартный тест Тьюринга. Тест Тьюринга дает определение разумного поведения. Его суть состоит в том, что система признается разумной, если общение с ней не позволяет отличить ее от человека по функциональным и интеллектуальным возможностям, то есть она умеет решать задачи и отвечать на вопросы.

Основной задачей естественных биологических систем является адаптация или просто выживание в широком смысле этого слова. Это выживание индивидуума, семьи и вида в целом. Главное содержание такого выживания состоит в сохранении информации, в первую очередь биологической (ДНК). Под разумным поведением мы будем понимать стратегии, которые максимально эффективно сохраняют, перерабатывают и трансформируют информацию, в соответствии со случайно меняющимися внешними условиями. Одной из важных задач организма является сохранение постоянства внутренней информации (метаболизм – обмен веществ), что существенно отличает естественные системы от искусственного интеллекта. Во многом задача разумного поведения состоит в таких реакциях организма, которые сохраняют неизменно его внутренние состояние.

Для решения задач искусственного интеллекта созданы языки искусственного интеллекта, к ним относятся в первую очередь LISP и PROLOG. Другим подходом к искусственному интеллекту являются

искусственные нейронные сети и генетические алгоритмы. Нейрон является элементарной клеткой нервной системы. Нейроны могут иметь различные элементы строения, но общая структура у них одинаковая.

1.2. Экспертные системы. Сочетание теоретического понимания проблемы и набора эвристических правил для ее решения, как показывает опыт, часто эффективны в конкретной предметной области. С помощью заимствования знаний человеческого эксперта и кодирования их в форму, которую компьютер может применить к аналогичным проблемам, создаются *экспертные системы*.

Стратегии экспертных систем основаны на знаниях человека-эксперта. Хотя многие программы пишутся самими носителями знаний о предметной области, большинство экспертных систем - результат сотрудничества между экспертом (врач, химик, геолог, строитель экономист) и независимым специалистом по искусственному интеллекту. *Эксперт* предоставляет необходимые знания предметной области, описывая свои методы принятия решений и демонстрируя эти навыки на тщательно отобранных примерах. Специалист по искусственному интеллекту, или инженер по знаниям, отвечает за реализацию этого знания в программе, которая должна работать эффективно и разумно. Экспертные способности программы проверяют, давая ей решать пробные задачи. Эксперт подвергает критике поведение программы, и в ее базу знаний вносятся необходимые изменения. Процесс повторяется, пока программа не достигнет требуемого уровня работоспособности.

Одной из первых систем, использовавших специфичные для предметной области знания, эта система DENDRAL, разработанная в Стэнфорде в конце 1960-х [Lindsay и др., 1980]. Она была задумана для определения строения органических молекул и химических молекул. Интересно отметить, что большинство экспертных систем были написаны для специализированных предметных областей. Эти области довольно хорошо изучены и располагают четко определенными стратегиями принятия решений. Проблемы,

определенные на нечеткой основе "здравого смысла", подобными средствами решить сложнее, поэтому не следует переоценивать возможности этой технологии.

Основные проблемы при создании экспертных систем:

1. Трудности в передаче "глубоких" знаний предметной области.
2. Недостаток здравомыслия и гибкости. Если людей поставить перед задачей, которую они не в состоянии решить немедленно, то они обычно исследуют сперва основные принципы и вырабатывают какую-то стратегию для подхода к проблеме Экспертным системам этой способности не хватает.
3. Неспособность предоставлять осмысленные объяснения. Экспертные системы не владеют глубоким знанием своей предметной области, и их пояснения обычно ограничиваются описанием шагов, которые система предприняла в поиске решения. Они зачастую не могут пояснить, "почему" был выбран конкретный подход.
4. Трудности в тестировании. Хотя обоснование корректности любой большой компьютерной системы достаточно трудоемко, экспертные системы проверять особенно тяжело. Это серьезная проблема, поскольку технологии экспертных систем применяются для таких критичных задач, как управление воздушным движением, ядерными реакторами и системами оружия.
5. Ограниченные возможности обучения на опыте. Сегодняшние экспертные системы делаются "вручную"; производительность разработанной системы не будет возрастать до следующего вмешательства программистов. Это заставляет серьезно усомниться в разумности таких систем.

Несмотря на эти ограничения, экспертные системы доказали свою ценность во многих важных приложениях.

1.3. Понимание естественных языков и семантическое моделирование. Одной из долгосрочных целей искусственного интеллекта

является создание программ, способных понимать человеческий языки строить фразы на нем. Способность применять и понимать естественный язык является фундаментальным аспектом человеческого интеллекта. Его успешная автоматизация привела бы к неизмеримо большей эффективности самих компьютеров. Многие усилия затрачены на написание программ, понимающих естественный язык. Хотя такие программы и достигли успеха в ограниченных случаях, натуральные языки с гибкостью и общностью, характерной для человеческой речи, лежат за пределами сегодняшних методологий.

1.4. Языки реализации искусственного интеллекта. Многие из применяемых методик сегодня являются стандартными методами разработки программного обеспечения и мало соотносятся с основами теории искусственного интеллекта. Другие же, такие объектно-ориентированное программирование, имеют значительный теоретический и практический интерес. Наконец, многие алгоритмы искусственного интеллекта сейчас реализуются на таких традиционных для вычислительной техники языках, как C++ и Java.

Языки, разработанные для программирования ИИ, тесно связаны с теоретической структурой этой области. Базовыми языками программирования искусственного интеллекта являются языки LISP, и PROLOG.

1.5. Машинное обучение. Обучение остается трудной проблемой искусственного интеллекта. Важность обучения несомненна, поскольку эта способность является одной из главных составляющих разумного поведения. Экспертная система может выполнять долгие поиски зашумленной информации, как, например, лицо в затемненной комнате или разговор на шумной вечеринке.

Более пригодны для сопоставления зашумленных и недостаточных данных *нейронные архитектуры*, поскольку они хранят знания в виде большого числа мелких элементов, распределенных по сети.

С помощью *генетических алгоритмов* и методик искусственной жизни вырабатываются новые решения проблем из компонентов предыдущих решений. Генетические операторы, такие как скрещивание или мутация, подобно своим эквивалентам в реальном мире, вырабатывают с каждым поколением все лучшие решения. В искусственной жизни новые поколения создаются на основе функции "качества" соседних элементов в прежних поколениях.

И нейронные архитектуры, и генетические алгоритмы дают естественные модели параллельной обработки данных, поскольку каждый нейрон или сегмент решения представляет собой независимый элемент.

Люди быстрее справляются с задачами, когда получают больше информации, в то время как компьютеры, наоборот, замедляют работу. Это замедление происходит за счет увеличения времени последовательного поиска в базе знаний. Архитектура с массовым параллелизмом, например человеческий мозг, не страдает таким недостатком. Наконец, есть нечто очень привлекательное в подходе к проблемам интеллекта с позиций нервной системы или генетики. В конце концов, мозг есть результат эволюции, он проявляет разумное поведение и делает это посредством нейронной архитектуры.

1.6. Искусственный интеллект и философия. Важно осознавать, что современный ИИ не только наследует богатую интеллектуальную традицию, но и делает свой вклад в нее. Например, поставленный Тьюрингом вопрос о разумности программ отражает наше понимание самой концепции разумности.

Что такое разумность, как ее описать?

Какова природа знания?

Можно ли его представить в устройствах?

Что такое навыки?

Может ли знание в прикладной области соотносится с навыком принятия решений в этой среде?

Как знание о том, что есть истина, соотносится со знанием как это сделать ("практика")?

Ответы на эти вопросы составляют важную часть работы исследователей и разработчиков ИИ. В научном смысле программы ИИ можно рассматривать как эксперименты. Проект имеет конкретную реализацию в виде программы, и программа выполняется как эксперимент. Разработчики программы изучают результаты, а затем перестраивают про граммы и вновь ставят эксперимент. Таким образом, можно определить, являются ли наши представления и алгоритмы достаточно хорошими моделями разумного поведения.

Ньюэлл и Саймон предложили более сильную модель интеллекта в гипотезе физической символьной системы: физическая система проявляет разумное поведение тогда и только тогда, когда она является физической символьной системой.

Многие применения искусственного интеллекта подняли глубокие философские вопросы. В каком смысле можно заявить, что компьютер "понимает" фразы естественного языка? Продуцирование и понимание языка требует толкования символов. Недостаточно правильно сформировать строку символов. Механизм понимания должен уметь приписывать им смысл или интерпретировать символы в зависимости от контекста.

Что такое смысл? Что такое интерпретация?

Подобные философские вопросы встают во многих областях применения искусственного интеллекта, от построения экспертных систем до разработки алгоритмов машинного обучения.

Сегодня известно также много важных биологических и социологических моделей обучения. Они будут рассмотрены в лекциях посвященных коннекционистскому и эмерджентному обучению. Успешность программ машинного обучения наводит на мысль о существовании универсальных принципов, открытие которых позволило бы конструировать программы, способные обучаться в реальных проблемных областях.

1.7. Нейронные сети и генетические алгоритмы. В ряде методик для реализации интеллекта используются явные представления знаний и тщательно спроектированные аалгоритмы перебора. Отличный подход состоит в построении интеллектуальных программ с использованием моделей, имитирующих структуры нейронов в человеческом мозге или эволюцию разных альтернативных конфигураций, как это делается в генетических алгоритмах и искусственной жизни.

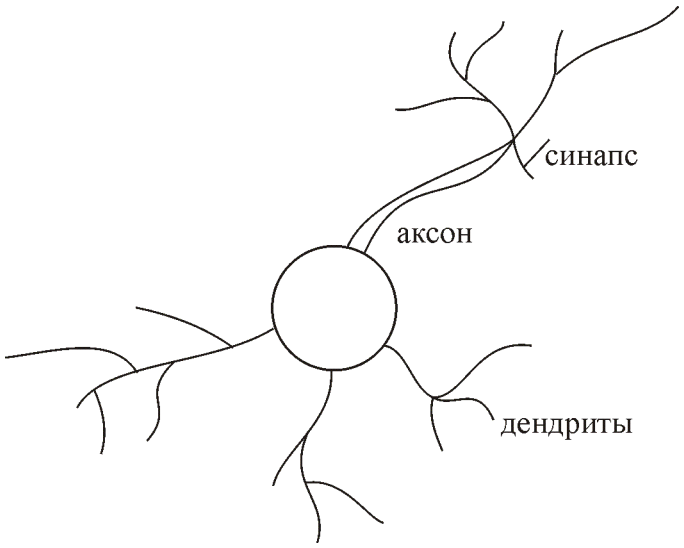


Рис.1.1. Типичный вид биологического нейрона

Схематически *нейрон* (рис.1.1) состоит из клетки, которая имеет множество разветвленных отростков, называемых *дендритами*, и одну ветвь — *аксон*. Дендриты принимают сигналы от других нейронов. Когда сумма этих импульсов превышает некоторую границу, нейрон сам возбуждается и импульс или "сигнал" проходит по аксону. Разветвления на конце аксона образуют синапсы с дендритами других нейронов. *Синапс* —это точка контакта между нейронами. Синапсы могут быть возбуждаюищми (excitatory) или тормозящими (inhibitory), в зависимости от того, увеличивают ли они результирующий сигнал.

По дендритам импульсы поступают в клетку. Дендриты суммируют воздействия, поступающие от других нейронов. Когда суммарное действие

суммарных импульсов превышают некоторый порог, тогда нейрон разряжается в виде набора импульсов, который называется *спайк*, а передавая возбуждение через аксон и синоптическую связь к другим нейронам.

Для того чтобы решать сложные и плохо формализуемые, задачи возникло направление, которое называется искусственные нейронные сети.

Искусственные нейронные сети состоят из нейроноподобных элементов, соединенных между собой в сеть. Существуют статические и динамические нейронные сети. В статических нейронных сетях изменение параметров системы происходит по некоторому алгоритму в процессе обучения. После обучения параметры сети не меняются. В динамических нейронных сетях отображение внешней информации, и ее обработка осуществляется в виде некоторого динамического процесса, т.е. процесса, зависящего от времени.

Естественный (биологический) интеллект возник в результате эволюции, основным инструментом которого является естественный отбор. Методы, основанные на аналогии с мутациями и естественного отбора в живых организмов, в теории искусственного интеллекта получили название генетического алгоритма.

1.8. Искусственный интеллект – некоторые выводы. ИИ - молодая и многообегающая область науки, основная цель которой —эффективный способ понимания и применении интеллектуального решения проблем, планирования и навыков общения к широкому кругу практических задач. Несмотря на разнообразие проблем, затрагиваемых исследователями ИИ, во всех отраслях этой сферы наблюдаются некоторые общие черты.

1. Использование компьютеров для доказательства теорем, распознавания образов, обучения и других форм рассуждений.

2. Внимание к проблемам, не поддающимся алгоритмическим решениям (эвристический поиск как основа методики решения задач в искусственном интеллекте).

3. Принятие решений на основе неточной, недостаточной или плохо определенной информации и применение формализмов представлений, помогающих справляться с этими проблемами.

4. Выделение значительных качественных характеристик ситуации.

5. Попытка решить вопросы семантического смысла и синтаксической формы.

6. Поиск решений, которые нельзя отнести к точным или оптимальным, но которые в каком-то смысле "достаточно хороши" на основе применения эвристических методов в ситуациях, когда получение оптимальных или точных решений невозможно.

Глава 2. Представление знаний

2.1. Теории смысла, основанные на ассоциациях. В основе искусственного интеллекта лежат способы представления информации. Существуют общие принципы организации знаний, которые применяются в различных областях и могут прямо поддерживаться соответствующим языком управления. Примеры иерархии классов можно найти как в научных, так и в общесмысловых, общеклассификационных системах.

Для представления знаний необходимо обеспечить общий механизм их представления, дать определения, сформулировать исключения, научить интеллектуальную систему делать предположения о недостающей информации. Возникает также проблема представления времени, причинности и неопределенности. Прогресс в создании интеллектуальных систем зависит от принципов организации знаний и их поддержки средствами представления высокого уровня.

В результате усилий психологов и лингвистов по оценке природы человеческого понимания возникли теории, основанные на ассоциациях. Эти теории определяют значения объекта в терминах сети ассоциации с другими объектами. С точки зрения такой теории восприятие объектов происходит через понятия.

Понятия – являются частью нашего знания о мире и связываются соответствующими ассоциациями с другими понятиями. Эти связи представляют свойства и поведения объектов. Например, понятие «снег» порождает ассоциативный ряд: холод, белый, скользкий, лед, снежный человек.

Попытка вспомнить отдельные свойства объекта зависит от последовательности ассоциаций. Последовательность ассоциаций определяет путь поиска в структуре памяти. Эффективным средством для формализации теории на основе ассоциации являются графы, за счет точного представления отношений посредством дуг и узлов. Семантическая сеть представляет знания в виде графа, узлы которого соответствуют фактам или понятиям, а дуги - отношениям или ассоциациям между понятиями.

Термин «семантическая сеть» обозначает семейство представлений, основанных на графах. Эти представления отличаются главным образом именами узлов, связи и выводами, которые можно делать в этих структурах. Общее множество предположений и отношений содержится во всех языках представления сетей. Одним из наиболее современных языков сетевого представления является концептуальный граф.

2.3. Стандартизации сетевых отношений. Само по себе представление отношений в виде графов имеет мало преимуществ перед исчислением предикатов. Сила сетевых представлений состоит в определении связей и специфических правил вывода, определяемых механизмом наследования.

Наиболее значительные особенности формализма семантических сетей состоят в наличии помеченных дуг и связей, и иерархических связей и выводов на основе ассоциативных связей. Одной из наиболее удачных попыток формального моделирования семантических структур естественного языка является теория концептуальной зависимости.

Синтаксис – формальная структура языка (грамматика).

Семантика – наука о смысловых значениях, понятиях и символах.

В рамках теории концептуальной зависимости рассматриваются 4 типа примитивов (базовых функций), на основе которых определяется смысл выражений. К ним относятся:

1) действия;

2) объекты;

3) модификаторы действий;

4) модификаторы объектов.

2.4. Сценарии. Программа, понимающая естественный язык, должна использовать большое количество исходных знаний, чтобы понять даже простейший разговор. Согласно экспериментальным данным, люди организуют знания в структуры, соответствующие типовым ситуациям. Читая статью о ресторанах или политике, мы устраняем двусмысленность в тексте с учетом тематики статьи. Если сюжет статьи неожиданно меняется, человеку необходима некая пауза для модификации структуры знания.

Слабо структурированный текст тяжело понять именно по тому, что мы не можем легко связать его с какой-либо из структур знаний. В этом случае ошибки понимания возникают из-за того, что мы не можем решить, какой контекст использовать для разрешения неопределенности местоимений и других двусмысленностей разговора.

Сценарий (script) – это структурированное представление, описывающее стереотипную последовательность событий в частном контексте.

Сценарии используются в системах понимания естественного языка для организации базы знаний в терминах ситуации, которые система должна понимать.

Сценарий включает следующие компоненты:

1. начальные условия, которые должны быть истинными при вызове сценария;
2. результаты или факты, которые являются истинными, когда сценарий завершен;
3. предположения, которые поддерживают контекст сценария;

4. роли, то есть действия, которые совершают отдельные участники;
5. сцены – этапы, разбивающие сценарий на временные последовательности.

Элементы сценария представляются отношением концептуальной зависимости. Собранные в одной структуре, они представляют собой последовательность значений или событий.

Пример: сценарий посещения ресторана.

1) Открытый ресторан; голодный посетитель с деньгами.
2) Клиент сыт; деньги потрачены; владелец ресторана получил прибыль.
3) Предположение: в ресторане есть столы и стулья, если не указано противоположное.
4) Официант принимает заказы, выставляет пищу, предоставляет счет. Клиент делает заказ, ест, платит.
5) Вход, заказ, принятие пищи.

Сценарий допускает рациональное предположение по умолчанию, которое является существенным в понимании языка.

Пример: Кто-то зашел в ресторан прошлым вечером. Он заказал бифштекс. Рассчитываясь, заметил, что остался без денег. Он поспешил домой, пока не начался дождь.

Подразумевается:

Он обедал (ужинал) прошлым вечером. Оплачивал, скорее всего, кредитной карточкой. Скорее всего, заказывал, используя меню.

2.5. Фреймы. *Фреймы* - это другая схема представления, подобная сценариям и ориентированная на включение строго организованной структуры данных, неявных информационных связей существующих в предметной области. Это представление поддерживает организацию знаний в более сложные единицы, которые отображают структуру объектов в этой области.

Суть теории фреймов в том, что при встрече с новой структурой из памяти выбирается структура, называемая фреймом. Этот каркас при

необходимости адаптируется и приводится в соответствии с реальным применением деталей. Фрейм может рассматриваться как статическая структура данных, используемая для представления хорошо понятных стереотипных ситуаций.

Например, достаточно один раз остановиться в гостинице, чтобы составить представление обо всех гостиничных номерах. Там имеется кровать, ванная, место для чемодана, телефон и др.

Каждый отдельный фрейм содержит следующую информацию:

1. Данные для идентификации фрейма;
2. Взаимосвязь этого фрейма с другими фреймами;
3. Дескрипторы требований (типичные диапазоны параметров);
4. Процедурная информация об использовании описанной структуры;
5. Информация по умолчанию (у стула 4 ножки, у телефона есть кнопки);
6. Информация для нового экземпляра. Содержит детали, которые могут быть уточнены для конкретной задачи.

Фреймы позволяют организовать иерархию знаний. Они расширяют возможности семантических сетей.

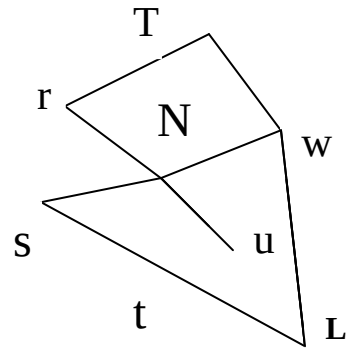
2.6. Концептуальные графы. Для моделирования семантики естественного языка и других областей был предложен ряд сетевых языков. Примером сетевого языка являются *концептуальные графы*.

Концептуальные графы – это конечный связный двудольный граф. Узлы графа представляют понятия или концептуальные отношения. В концептуальных графах метки дуг не используются. Отношения между понятиями представляются узлами концептуальных отношений. В концептуальных графах узлы понятий представляют собой либо абстрактные, либо конкретные объекты в мире рассуждений.

Иерархия типов. *Иерархия типов* - это частичное упорядочение на множестве типов, которое можно обозначить при помощи символов (≤).

s и *t* *t*<=*s* *t* – подтип *s*

s – супертип *t*



Если *s*, *t* и *u* – и типы, *t*≤*s*, *t*≤*u* , то говорят, что *t* – это общий подтип для s и *u*.

Подобным образом, если *s*≤*w* и *u*≤*w*, то *w* – общий супертип для *s* и *u*.

Иерархия типов в концептуальных графах представляет собой алгебраическую структуру, которая называется *решеткой*, описывающей общий вид системы множественного наследования. В этой решетке типы могут иметь множество родителей и детей. Однако каждая пара типов должна иметь минимальный общий супертип и максимально общий подтип.

Обобщение и специализация. Теория концептуальных графов включает операции для создания новых графов на основе существующих; они позволяют генерировать новые графы путем либо специализации, либо путем обобщения существующего графа. Существует 4 важнейшие операции:

- копирование;
- ограничение;
- объединение;
- упрощение.

Копирование создает точную копию графа.

Ограничение позволяет заменить узлы и понятия графа узлами, представляющими их специализацию.

Объединение. Правило объединения позволяет интегрировать два графа в один.

Упрощение позволяет при наличии двух одинаковых отношения вычеркнуть одно из них.

Эти правила не являются правилами вывода. Они не гарантируют, что из истинных графов будут выводиться истинные графы.

Концептуальные графы и логика. Хотя концептуальные графы можно описать с помощью исчисления предикатов, они поддерживают ряд специальных механизмов вывода, таких как объединение и ограничение, не являющихся частью исчисления предикатов.

Глава 3. Нейрофизиологические данные об обработки информации в биологических системах

3.1. Физиологические основы мышления. Мышление мы связываем только с деятельностью материальной нервной системы, функционированием мозга. Основой функционирования мозга является работа нейронных клеток, объединенных в гигантскую нейронную сеть. Внутренние механизмы функционирования нейронных клеток не имеют прямого отношения к феномену мышления, а важны лишь функциональные отклики нейронов на электрические и химические сигналы.

Нейроны являются макроскопическими образованиями, и их работа в нейронных сетях не связана с проявлением квантовых явлений, то есть может рассматриваться в рамках классических представлений. Мышление может быть понято, как сложная коллективная динамика гигантских нейронных сетей. Мышление функционирует по своим особым законам, которые непосредственно связаны со структурой клеточных связей мозга. Раскрытие связи различных свойств и феноменов мышления со структурой и динамикой нейронных сетей и означает понимание работы мозга сущности мышления.

3.2. Работа нейронов и нейронных цепей. Реакция нейронов на внешние сигналы поступающих по дендритам проявляются через так

называемый потенциал действия. Одним из основных свойств потенциала действия является то, что это взрывное, пороговое событие, возникающее по законам «все, или ничего». Отклик клетки происходит в виде импульса или серии импульсов, в так называемых спайках.

Нейроны организованы в *нейронные цепи*, выполняющие различные функции. Техника освещения определенных зон сетчатки позволила выделить *рецептивные поля* в *коре человеческого мозга*. Рецептивные поля нейрона зрительной коры – это зона сетчатки, при попадании света на которую, может изменяться активность данной нейронной коры. Нейроны, обрабатывающие сходную информацию расположены близко друг к другу, поскольку их рецептивные поля сильно перекрываются. Это позволяет таким нейронам легко взаимодействовать друг с другом. Зрительная информация поступает в кору, расположенную в затылочной части.

Для всей коры мозга характерно шестислойное строение. Развитие нервной системы сопровождается ее ростом и установлением связей между нейронами. В дальнейшем под воздействием внешних условий и поступающей информации часть связей между нейронами усиливается, а часть связей – ослабевает. Большое количество нейронов гибнет, что также является одним из механизмов формирования нервной системы.

3.3. Строение коры человеческого мозга. В коре мозга выделяют разные структуры, из которых функция сознания связана с так называемой новой корой головного мозга. Кора состоит из шести слоев клеток. В вертикальном направлении клетки коры организованы в *колонки нейронов*. Каждая колонка насчитывает около 100 нейронов. Соседние колонки слабо связаны друг с другом в нижних слоях. Только нижние слои коры связаны непосредственно с периферией. В верхних слоях нейроны сильно связаны друг с другом, что позволяет выполнять сложные интегрирующие и мыслительные функции. Расположение нейронов в виде колонок делает возможным картирование одновременно нескольких переменных на двумерной матрице поверхности коры.

3.4. Сознательная деятельность. Отдельные системы коры головного мозга имеют иерархическое строение. Возбуждение, возникающее в периферических органах, чувств сначала приходит в первичные или проекционные зоны. Отдельные участки этих зон представляют собой расположенные по топографическому принципу проекции соответствующих периферических рецепторов. Затем, возбуждение распространяется на вторичные зоны коры, которые на основе работы верхних ассоциативных слоев нейронов выполняют интегрирующие функции. Они объединяют топографические проекции, возникшие не периферии возбуждения в сложной, функционально-организованной системе.

Важнейшую роль в процессах обработки информации мозга играет внимание, представляющее сосредоточенность деятельности субъекта в данный момент времени на каком-то реальном или идеальном объекте (предмете, событии, суждении). Важную роль в работе мозга играют и модулирующие системы. Эти системы не выполняют непосредственно обработку информации, а влияют на уровень активности тех или иных структур в системе мозга.

3.5. Афферентный синтез и теория функциональных систем. Автором теории функциональных систем является Анохин. *Функциональной системой* называется такой комплекс избирательно-вовлеченных компонентов, у которого взаимодействия и взаимоотношения принимают характер взаимосодействия компонентов для получения фокусированного конечного полезного результата. Функциональный принцип выборочной мобилизации структур является доминирующим в функциональных системах.

При образовании иерархии систем всякий более низкий уровень системы должен организовать контакт результатов, что и может составить следующий более высокий уровень системы. В этом случае иерархия систем превращается в иерархию результатов каждой из субсистем предыдущего уровня.

Поведенческий акт любой сложности начинается со стадии афферентного синтеза. Возбуждение, вызванное внешним стимулом, действует не изолированно. Оно непременно вступает во взаимодействие с другими афферентными возбуждениями, имеющими иной функциональный смысл. Мозг непрерывно обрабатывает все сигналы, поступающие по многочисленным сенсорным каналам. И только в результате синтеза этих афферентных возбуждений создаются условия для реализации определенного целенаправленного поведения. Содержание афферентного синтеза определяется влиянием нескольких факторов: мотивационного возбуждения, памяти, обстановочной и пусковой афферентации.

Глава 4. Идея и реализация искусственного нейрона

4.1. Искусственный нейрон. В основе искусственной нейронной сети лежит отдельный элемент – *искусственный нейрон*. Математическая модель нейрона представляет собой абстрактный элемент, который имеет несколько входов и один выход. Структурная модель искусственного нейрона показана на рис. 4.1.

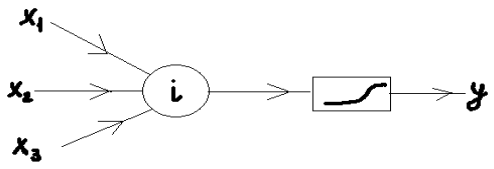


Рис. 4.1. Структурная модель искусственного нейрона

На вход *i*-ого нейрона подается набор *N* сигналов $x_j, j = 1,2,\text{K}\ N$. Далее эти сигналы суммируются с *весами* w_{ij} , и к результату прибавляется постоянное значение b_i. Получается некоторая величина u_i:

$$u_i = \sum_{i=1}^{N} w_{ij} x_j + b_i \,, \qquad (4.1)$$

где b_i - смещение.

Для получения y_i на выходе вычисляется *функция активации* $f(u_i)$ или *передаточная функция*. На выходе нейрона получается значение:

$$y_i = f(u_i). \qquad (4.2)$$

Нами здесь описан статический искусственный нейрон.

4.2. Виды функции активации. Наиболее характерный вид функции активации или функции отклика – сигмаидальный (похожий на «хвостик» греческой буквы σ), - он показан на рис.3.2.

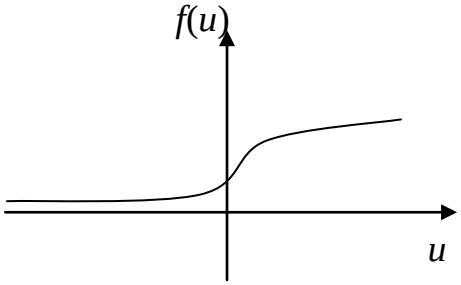


Рис. 4.2. Сигмаидальная функция отклика нейрона

Часто используются также *радиальные базисные функции*, отклик которых имеет зависимость от аргумента, показанную на рис. 4.3.

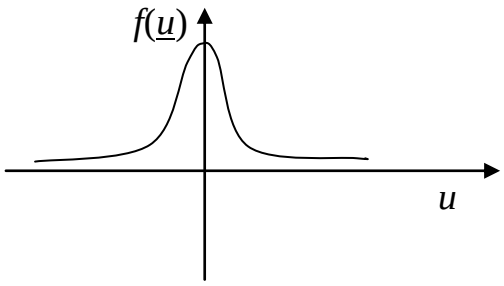
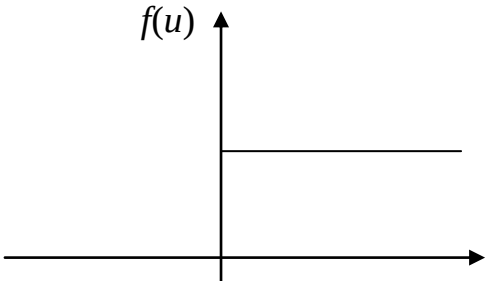


Рис. 4.3. Функция отклика нейрона для радиальной базисной функции

В простейшем случае функция активации имеет вид ступеньки (рис. 4.4):



1

u

Рис. 4.4. Ступенчатая функция отклика

$$f(u) = \begin{cases} 1, u \geq 0 \\ 0, u < 0 \end{cases}, \tag{4.3}$$

$f(u) = \theta(u)$ - тета–функция. Нейрон со ступенчатой функцией активации называется *персептроном Мак Каллока–Питса*.

Для нейрона с сигмаидальной функцией активации часто применяется зависимость в виде униполярной функции:

$$f(u) = \frac{1}{e^{-\beta u} + 1}, \beta > 0. \tag{4}$$

При $u \to -\infty$, $e^{-\beta u} \to \infty$ и $f(u) \to 0$, а при $u \to +\infty$, $e^{-\beta u} \to 0$ и $f(u) \to 1$.

Часто используют также биполярную функцию:

$$f(u) = \frac{e^{\beta u} - e^{-\beta u}}{e^{\beta u} + e^{-\beta u}}, \beta > 0. \tag{4.5}$$

При $u \to -\infty$, $f(u) \to -1$, а при $u \to +\infty$, $f(u) \to 1$. Чаще всего на практике полагают β=1. Важным свойством сигмаидальной функции является её дифференцируемость и простое выражение для производной.

Для униполярной функции:

$$\frac{df(u)}{du} = \beta \cdot f(u) \cdot [1 - f(u)]. \tag{4.6}$$

Для биполярной функции:

$$\frac{df(u)}{du} = \beta \cdot [1 - f(u)]. \tag{4.7}$$

Для радиальных базисных функций часто используют функцию Гаусса:

$$f(u) = \exp(-u^2 / 2) \tag{4.8}$$

с производной

$$\frac{df(u)}{du} = -uf(u). \qquad (4.9)$$

При $u \to \pm\infty$ $u^2 \to \infty$, и $f(u) \to 0$. При u =0 значение $f(0)=1$.

Искусственные нейроны должны воспроизводить сигналы, лежащие в самых разных диапазонах. Однако стандартный нейрон может выдавать значение в интервале [0,1] или [–1,1]. Для приведения в соответствие моделируемых сигналов и сигналов нейрона данные масштабируются, то есть делается замена $y \to y/\|y\|_{\max}$, и тогда $|y| \le 1$. Теперь при построении нейронной сети все значения функции не выходят за пределы [0, 1] или [-1, 1] сооветственно типу выбранной функции отклика.

4.3. Нейронные сети. Разработчиками *нейронных сетей* предполагается, что все особенности поведения нейронных сетей заключаются в их структуре. Такая точка зрения называется – коннекционизмом. *Коннекционизм* – это подход, согласно которому можно получить отображение любой структуры данных, используя для этого соответствующее соединение однотипных стандартных нейронов.

Фактически, для решения большинства простых практических задач достаточно трёхслойных нейронных сетей. Первый слой нейронов сети называется входным, последний слой – выходным. Нейроны, не имеющие прямых связей с входами и выходами сети, называются – внутренними (внутренние слои нейронной сети). На вход сети подаются некоторые известные сигналы, а на выходе ожидают получения известного выходного сигнала. Настройка сети заключается в выборе весов w_{ij} и смещений b_i. Эти параметры представляют собой некоторые числа, которые нужно подобрать так, чтобы на известных примерах система давала правильные ответы, в виде значений функции на выходе. При этом система модифицирует свои параметры по мере предъявления ей новых примеров с тем, чтобы наиболее точно воспроизводить выходной сигнал. Такое обучение

системы называется обучением *с учителем*. Возможно также *обучение без учителя*, когда нейросети просто предъявляются разные примеры данных. Мы остановимся главным образом на обучении с учителем.

Представимость произвольных непрерывных отображений на основе трёхслойной нейросети является следствием теоремы Колмогорова-Арнольда о представимости функции многих переменных в виде суперпозиции и сумм функции одной переменной.

4.4. Представимость данных и отображения. Пусть в систему поступают данные $\mathbf{x} = (x_1, x_2, \text{K}\ x_N)$. Для того, чтобы взаимно однозначно представить N величин, нужно столько же величин $\mathbf{x}' = (x_1', x_2', \text{K}\ x_N')$ внутри системы, и тогда, между векторами $\mathbf{x}$ и $\mathbf{x}'$ будет соответствие.

Построим отображение: $x_1, x_2 \to x_1'$. Это обычная функция двух переменных, то есть две переменные выражены через третью: $x_1' = F(x_1, x_2)$ Это уравнение задает поверхность в трехмерном пространстве.

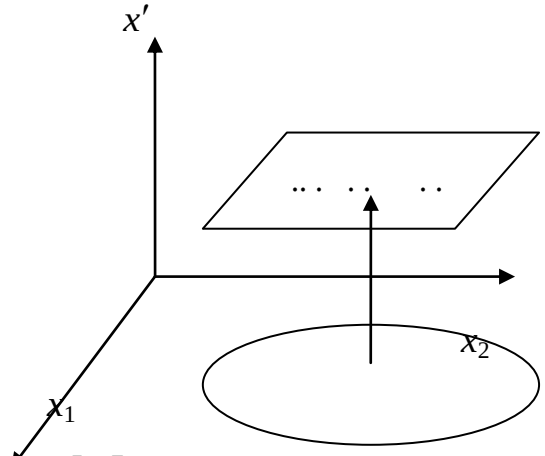


Рис. 4.5. Функция двух переменных

Если рассмотреть пересечение с этой поверхностью плоскости $x_1' = const$, то пересечением будет линия, а если поверхность сложная, то несколько линий. Значит, взаимно обратного отображения может и не быть.

Важным является вопрос о том, насколько хорошо можно аппроксимировать произвольную непрерывную функцию с помощью нейросети. Речь идет о представлении функции с произвольно заранее

заданной точностью. Пусть есть функция многих переменных: $f(x_1, x_2,..,x_n)$. Имеется теорема, в которой утверждается, что если мы возьмем произвольную монотонно растущую функцию $\varphi(u)$ (рис.3.6), то достаточно точно приближенное выражение можно записать в виде:

$$F(x_1, x_2, \text{K } x_n) = \sum_{i=1}^{m} a_i \varphi\left(\sum_{j=1}^{n} w_{ij} x_j + b_j \right). \qquad (4.10)$$

Это по существу *двухслойная нейросеть*, где φ ($\sum w_{ij} x_j + b_i$) – искусственный нейрон. Причем для всех $\varepsilon \to 0 \quad |f - F| < \varepsilon$.

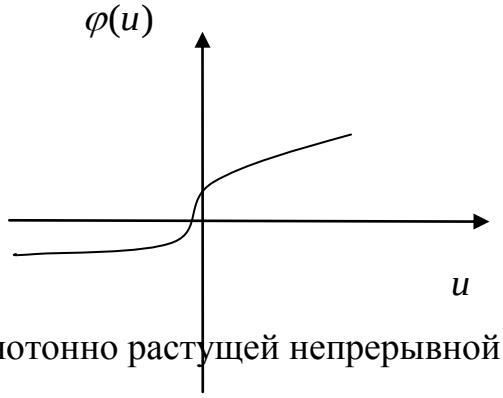


Рис. 4.6. Вид монотонно растущей непрерывной функции

Таким образом, теорема доказывает, что двухслойная нейросеть (рис. 4.7) является универсальным аппроксиматором, то есть позволяет воспроизвести любую функцию любой сложности.

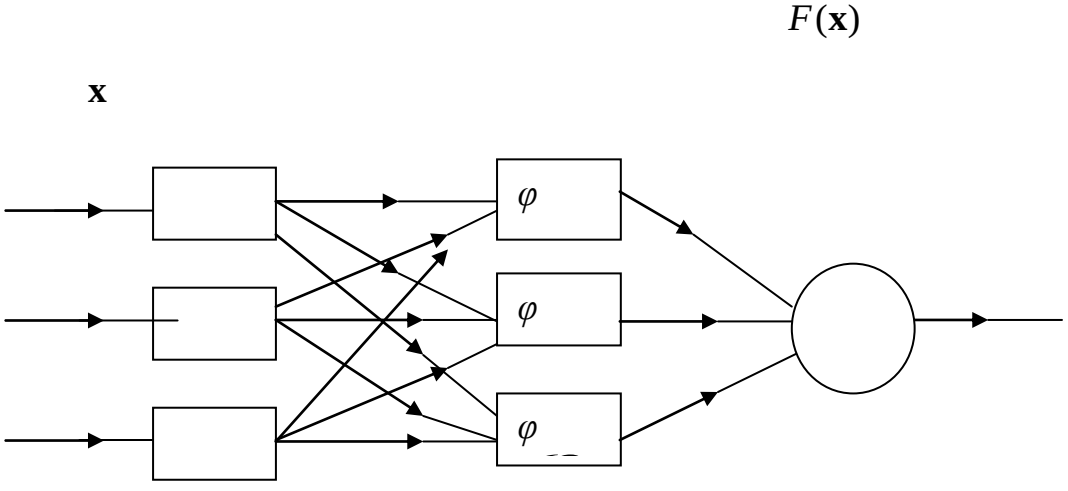


Рис. 4.7. Двухслойная нейронная сеть как универсальный аппроксиматор

Глава 5. Многослойный персептрон

5.1. Линейный персептрон. Если функция активации нейронов $f(u) = u$, то такой нейрон называется линейным (рис. 5.1).

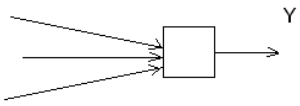


Рис. 5.1. Линейный персептрон

Иначе говоря,

$$y = \sum_{i=1}^{N} w_i x_i + b\,. \qquad (5.1)$$

При наличии одного входа *линейный персептрон* задаёт прямую:

$$y = ax + b. \qquad (5.2)$$

Если функциональная зависимость, которую мы хотим описать и есть прямая, то задача обучения нейрона заключается в определении коэффициентов a и b таким образом, чтобы отклонение прямой от заданных точек (рис. 5.2) было минимально.

Обозначим значение функции в точках как функцию $d = d(x)$

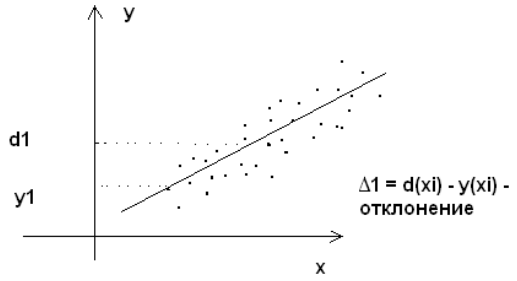


Рис. 5.2. Значения функции $d = d(x)$ в точках и проведенная по ним прямая

Чтобы уменьшить отклонение Δ проведенной кривой от точек данных (рис. 5.2), нужно ввести критерий меры отклонения. Разумнее всего для этого брать какое-то число, отражающее степень близости точек. Чтобы ввести соответствующее понятие, нужно определить расстояние между точкой на

прямой и точкой, представляющей данные, т.е. задать, например, евклидову метрику с помощью теоремы Пифагора.

Важнейшим для реализации нейронных сетей является определение алгоритма обучения сети. Хотя основные идеи нейронных сетей были развиты еще Розенблютом в конце сороковых годов, отсутствие эффективных методов обучения нейронных сетей тормозило их развитие и применение.

5.2. Алгоритм обратного распространения ошибки. В настоящее время одним из самых эффективных и обоснованных методов облучения нейронных сетей является *алгоритм обратного распространения ошибки*, который применим к *однонаправленным многослойным сетям*. В многослойных нейронных сетях имеется множество скрытых нейронов, входы и выходы которых не являются входами и выходами нейронной сети, а соединяют нейроны внутри сети, то есть *скрытые нейроны*.

Занумеруем выходы нейронной сети индексом $j = 1,2,\text{K},n$, а обучающие примеры индексом $M = 1,2,\text{K},M_0$. Тогда в качестве целевой функции можно выбрать функцию ошибки как сумму квадратов расстояний между реальными выходными состояниями y_{jM} нейронной сети, выдаваемых сетью на входных данных примеров, и правильными значениями функции d_{jM}, соответствующими этим примерам. Пусть $\mathbf{x} = \{x_i\}$ – столбец входных значений, где i=1,2,..,n. Тогда $\mathbf{y} = \{y_j\}$ – выходные значения, где j=1,2,…,m. В общем случае $n \neq m$. Рассмотрим разность $y_{jM} - d_{jM}$, где d_{ji} – точное (правильное) значение из примера. Эта разность должна быть минимальна. Введем расстояния согласно евклидовой метрике, определив норму

$$\|\mathbf{y} - \mathbf{d}\| = \sqrt{(\mathbf{y} - \mathbf{d}, \mathbf{y} - \mathbf{d})^2} \ . \qquad (5.3)$$

Пусть целевая функция имеет вид

$$E = \frac{1}{2} \sum_{j,M} (y_{j,M} - d_{j,M})^2 \ . \qquad (5.4)$$

Коэффициент ½ выбран из соображений более короткой записи последующих формул. Задача обучения нейронной сети состоит в том, чтобы найти такие коэффициенты $w_{\beta k}$, при которых достигается минимум $E(\mathbf{w})(E \geq 0)$. Обозначим q – номер произвольного внутреннего слоя нейронной сети, Q – номер выходного слоя.

Для простоты индексации рассмотрим случай, когда на вход подается только один пример, и целевая функция

$$E = \frac{1}{2}\sum_j (d_j - y_j)^2 . \qquad (5.5)$$

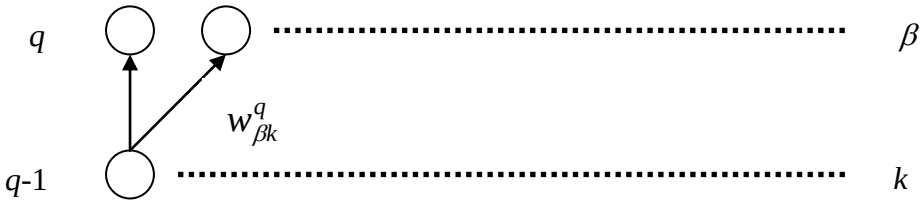


Рис. 5.3. Схема связей между нейронами слоя q и слоя $q+1$ в сети прямого распространения

Приращение целевой функции при малом изменении параметров сети в результате суммирования вклада нейронов в каждом слое q составит

$$dE = \sum_{q\beta k} \frac{\partial E}{\partial w^q_{\beta k}} \cdot dw^q_{\beta k} . \qquad (5.6)$$

Индекс q обеспечивает суммирование изменений по слоям, индекс β обеспечивает суммирование по нейронам слоя с номером q, а индекс k обеспечивает суммирование по выходам нейронов из слоя с номером $q-1$. Для движения в процессе обучения к минимуму величины E необходимы такие изменения весов $\mathbf{w}^q$, чтобы было обеспечено изменение целевой функции $dE<0$. Подходящим является выбор

$$dw_{\beta k}^{q} = -\eta \cdot \frac{\partial E}{\partial w_{\beta k}^{q}}, \qquad (5.7)$$

где малый параметр величины шага η>0 гарантирует такое поведение. Тогда

$$dE = -\sum_{q\beta k}\left(\frac{\partial E}{\partial w_{\beta k}^{q}}\right)^{2} \cdot \eta < 0, \qquad (5.8)$$

что обеспечивает уменьшение E при достаточно малом значении η. Из выражения (5.8) следует, что вклады каждого слоя, каждого нейрона и каждой связи нейрона в уменьшение величины E в этом случае независимы и аддитивны. Выходное значение каждого слоя определяется универсальной функцией активации.

Рассмотрим зависимость ошибки на выходе сети от параметров-весов в слое с номером q для нейрона с номером β и связи между нейроном предыдущего слоя с номером k. Вычислим соответствующую частную производную

$$\frac{\partial E}{\partial w_{\beta k}^{q}} = \left(\frac{\partial E}{\partial y_{\beta}^{q}} \cdot \frac{\partial y_{\beta}^{q}}{\partial u_{\beta}^{q}}\right) \cdot \frac{\partial u_{\beta}^{q}}{\partial w_{\beta k}^{q}}, \qquad (5.9)$$

где $\frac{\partial y_{\beta}^{q}}{\partial u_{\beta}^{q}} = \frac{df(u_{\beta}^{q})}{du_{\beta}^{q}}$, $\frac{\partial E}{\partial y_{\beta}^{q}} \cdot \frac{\partial y_{\beta}^{q}}{\partial u_{\beta}^{q}} = \delta_{\beta}^{q}$, а в выходном слое $q = Q$. В выходном слое со значениями на выходе y_i имеем

$$\frac{\partial E}{\partial y_k} = \frac{\partial}{\partial y_k}\left(\frac{1}{2} \cdot \sum_i (d_i - y_i)^2\right) = \frac{\partial}{\partial y_k} \cdot \frac{1}{2} \cdot (y_k - d_k)^2 = y_k - d_k. \qquad (5.10)$$

В произвольном текущем слое производная $\frac{\partial u_{\beta}^{q}}{\partial w_{\beta k}^{q}} = y_k^{q-1}$, где q – номер текущего слоя, $q-1$ - номер предыдущего слоя. Поскольку $u_{\beta}^{q} = \sum w_{\beta k}^{q} y_k^{q-1}$, то $\frac{\partial E}{\partial w_{\beta k}^{q}} = \delta_{\beta}^{q} y_k^{q-1}$.

Рассмотрим изменение параметров отклика сети, перейдя к переменным слоя, отстоящего на один слой дальше от выходного слоя:

$$\frac{\partial E}{\partial y_\alpha^q} = \sum_k \left(\frac{\partial E}{\partial y_\alpha^{q+1}} \cdot \frac{\partial y_\alpha^{q+1}}{\partial u_k^{q+1}} \right) \cdot \frac{\partial u_k^{q+1}}{\partial y_\alpha^q}. \tag{5.11}$$

Умножим обе части на $\frac{\partial y_\alpha^q}{\partial u_\alpha^q}$:

$$\frac{\partial y_\alpha^q}{\partial u_\alpha^q} \cdot \frac{\partial E}{\partial y_\alpha^q} = \sum_k \left(\frac{\partial E}{\partial y_\alpha^{q+1}} \cdot \frac{\partial y_\alpha^{q+1}}{\partial u_k^{q+1}} \right) \cdot \frac{\partial u_k^{q+1}}{\partial y_\alpha^q} \cdot \frac{\partial y_\alpha^q}{\partial u_\alpha^q}, \tag{5.12}$$

и далее обозначим

$$\delta_k^{q+1} = \frac{\partial E}{\partial y_\alpha^{q+1}} \cdot \frac{\partial y_\alpha^{q+1}}{\partial u_k^{q+1}}, \tag{5.13}$$

а

$$\delta_k^q = \frac{\partial y_\alpha^q}{\partial u_\alpha^q} \cdot \frac{\partial E}{\partial y_\alpha^q}. \tag{5.14}$$

Тогда получим

$$\delta_j^q = \left[\sum_k \delta_k^{q+1} \cdot \frac{\partial u_k^{q+1}}{\partial y_\alpha^q} \right] \cdot \frac{dy_\alpha^q}{du_\alpha^q}. \tag{5.15}$$

Для униполярной функции активации

$$f(u) = \frac{1}{1+e^{-u}}, \frac{df}{du} = f(u) \cdot (1 - f(u)), \tag{5.16}$$

и параметры вычисляются в соответствии с формулами

$$\delta_j^Q = [f_j^Q (1 - f_j^Q)] \cdot (y_j^Q - f_j^Q); \; \delta_j^q = [f_j^q (1 - f_j^q)] \cdot \sum_k \delta_k^{q+1} w_{jk}^{q+1}, \tag{5.17}$$

где y_j^Q - правильное значение на выходе последнего слоя сети. При этом

$$\Delta w_{\beta k}^q = -\eta \cdot \delta_\beta^q \cdot y_k^{q-1}. \tag{5.18}$$

5.3. Алгоритм вычисления весов нейронов в соответствии с методом обратного распространения ошибки:

Шаг 1. Подать на входы сети один из примеров и вычислить все значения в сети от входа к выходу.

Шаг 2. Рассчитать δ_j^Q.

Шаг 3. Рассчитать δ_j^q и Δw_{ik}^q.

Шаг 4. Вычислить скорректированные веса нейронов

$$w_{ij}^q(t) = w_{ij}^q(t-1) + \Delta w_{ij}^q(t),$$

где t – номер шага.

Шаг 5. Если ошибка сети существенна (мы сравниваем контрольный результат и то, что получили), то снова переходим к шагу 1. Иначе обучение прекращается.

Для лучшей сходимости алгоритма предпочтительно обучающие примеры подавать вразбивку.

Формулы, задающие порядок вычислений в соответствии с описанным алгоритмом имеют вид:

$$f(u) = \frac{1}{1+e^{-u}}, \tag{5.19}$$

Для выходного слоя вычисляются вспомогательные величины

$$\delta_j^Q = [f_j^Q(1-f_j^Q)]\cdot(y_j^Q - f_j^Q), \tag{5.20}$$

где j – номер нейрона в выходном слое. В последующих слоях

$$\delta_j^q = [f_j^q(1-f_j^q)]\cdot\sum_k \delta_k^{q+1} w_{jk}^{q+1}, \tag{5.21}$$

$$\Delta w_{\beta k}^q = -\eta\cdot\delta_\beta^q\cdot y_k^{q-1}. \tag{5.22}$$

Для двухслойной сети с тремя нейронами, показанной на рис.4:

$$\delta^2 = (f_2 - d_2)\cdot f_2\cdot(1-f_2), \tag{5.23}$$

$$\delta_1^1 = \delta^2\cdot f_1^1\cdot(1-f_1^1)\cdot w_1^2, \tag{5.24}$$

$$\delta_2^1 = \delta^2\cdot f_2^1\cdot(1-f_2^1)\cdot w_2^2. \tag{5.25}$$

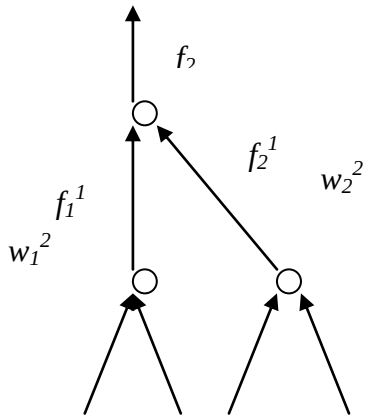


Рис. 5.4. Двухслойная нейронная сеть прямого распространения

Искусственные нейронные сети прямого распространения весьма популярны в практических приложениях, поскольку имеют простую и наглядную структуру, способны решать довольно сложные задачи и имеют простые и эффективные алгоритмы обучения, в первую очередь алгоритм обратного распространения ошибки.

5.4. Градиентные методы обучения и метод наискорейшего спуска. Метод обратного распространения ошибки дает частный способ поиска минимума целевой функции обучения. В целом обучение мажет рассматриваться как задача о наикратчайшем спуске в самую глубокую долину рельефа целевой функции, зависящего от вектора весов сети $\mathbf{w}$.

Рассмотрим в начале простой двумерный случай, то есть пусть существуют веса w_1, w_2 , а отклонение от цели $E=1/2(y-d)^2$ **.** Для наискорейшего спуска нужно идти по линии максимальной крутизны (для одной переменной - по касательной). Для этого требуется построить перпендикуляр к линии уровня, то есть вычислить градиент функции E. Для метаматематической формулировки этого процесса разложим целевую функцию в ряд Тейлора в окрестности некоторой текущей точки:

$$E = E_0 + \sum_{\alpha k q} \frac{\partial E}{\partial w_{\alpha k}^q} \Delta w_{\alpha k}^q + \frac{1}{2!} \sum\sum \frac{\partial^2 E}{\partial w_{\alpha n}^q \cdot \partial w_{\beta l}^m} \cdot \Delta w_{\alpha n}^q \cdot \Delta w_{\beta l}^m + O(\Delta w)\text{K} \quad .(5.26)$$

Введем единый индекс весов $i = \{q, \alpha, n\}$. Тогда

$$\frac{\partial^2 E}{\partial w_i \cdot \partial w_j} = \gamma_{ij}, \tag{5.27}$$

$$\gamma_{ij} = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1 \cdot \partial w_2} & \Lambda & \Lambda \\ \Lambda & \Lambda & \Lambda \\ \Lambda & \Lambda & \frac{\partial^2 E}{\partial w_i \cdot \partial w_j} \end{bmatrix}. \tag{5.28}$$

В точке экстремума все частные производные =0, а $\det \gamma_{ij}$ может быть любым по знаку:

1) $\det \gamma_{ij}$=0 – это точка перегиба (седло или перевал).

2) $\det \gamma_{ij}$>0 – точка минимума,

3) $\det \gamma_{ij}$<0 – точка максимума.

Для построения алгоритма обучения возьмем разложение функции в достаточно малой окрестности текущего значения весов нейронов сети, а остальные слагаемые ряда отбросим. В простом градиентном методе мы отбрасываем все, кроме первой поправки к значению функции:

$$E = E_0 + \sum_{\alpha k q} \frac{\partial E}{\partial w_{\alpha k}^q} \Delta w_{\alpha k}^q. \tag{5.29}$$

Дадим приращение весам

$$\Delta w_{\beta k}^q \sim -\eta \cdot \frac{\partial E}{\partial w_{\alpha k}^q}. \tag{5.30}$$

Здесь $\frac{\partial E}{\partial w_i} = (\mathrm{grad} E)_i$, то есть $\frac{\partial E}{\partial w_i}$ является компонентой градиента.

$$(\nabla E)_i = \left(\frac{\partial E}{\partial w_i} \right). \tag{5.31}$$

Отличие от метода обратного распространения ошибки состоит в том, что раньше производные вычислялись аналитически, а теперь целиком численно:

то есть даем приращение Δw_i и находим: $\frac{\partial E}{\partial w_i} \approx \frac{\Delta E}{\Delta w_i}$. Это один из самых распространенных методов, основанный на определении величины производной.

Отметим трудности данного метода обучения сети:

1. Если использовать большой шаг при спуске, то будет происходить колебание вокруг точки минимума, как показано на рис. 5.5.

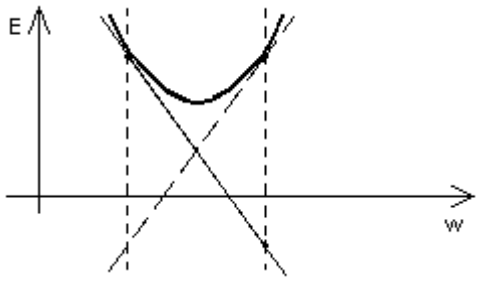


Рис. 5.5. Иллюстрация плохой сходимости градиентного спуска при большом шаге в области минимума

2. В точке перегиба производные $\frac{\partial E}{\partial w_i}$=0, поэтому обучение прекратится, не достигнув минимума. Чтобы этого избежать, используют метод моментов. Для прохождения области перегиба необходимо придать обучению некоторую инерцию, поэтому вводится дополнительная к (30) поправка

$$\alpha\left(w_{\beta j}^{q}(t) - w_{\beta j}^{q}(t-1)\right)$$

к величине $w_{\beta j}^{q}$, а параметр α должен быть порядка 4%. Это позволяет проскочить «плато», то есть область медленного изменения целевой функции, в процессе обучения.

3. Проблема локального минимума. При обучении градиентным методом сеть может попасть в состояние локального минимума E (рис. 5.6).

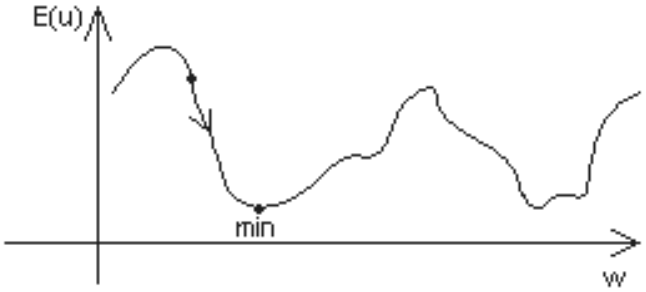


Рис. 5.6. Пример функциональной зависимости с несколькими локальными минимумами

При попадании в локальный минимум для выхода из него нужно осуществить старт алгоритма с новых начальных значений весов сети. Для их задания можно использовать, в частности, алгоритм случайного блуждания в пространстве весов $\mathbf{w}$.

Глава 6. Сети Хемминга

6.1. Распознающая сеть. В целом ряде задач не требуется воспроизводить распознаваемый образ, как это делают сети Хопфилда, а достаточно указать номер эталона, ближайшего к предъявленному входному вектору. Для этого может быть использована сеть Хемминга. Преимуществами этой сети по сравнению с сетью Хопфилда являются меньшие затраты на память и объем вычислений.

Нейронная сеть Хемминга показана на рис. 6.1 и состоит из трех слоев: входного, скрытого и выходного. Скрытый и выходной слои содержат по k нейронов, где k - это число эталонных образов.

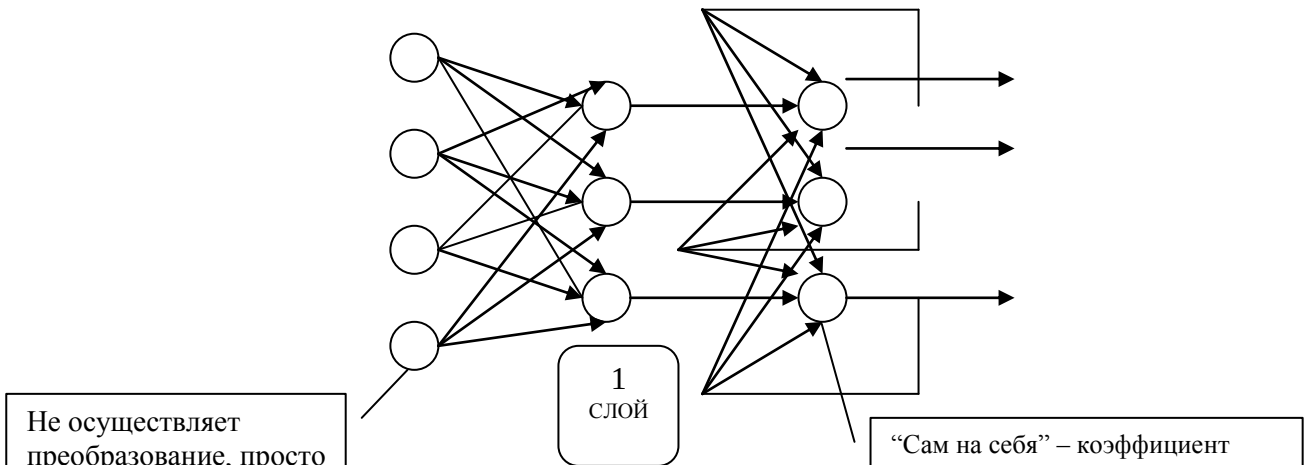


Рис. 6.1. Схема сети Хемминга

Каждый из нейронов скрытого слоя соединен с выходами *n*-нейронов входного слоя. Выходы нейронов выходного слоя связаны со входами остальных нейронов этого слоя отрицательными обратными (ингибиторными или тормозящими) связями. Единственная положительная обратная связь подается с выхода для каждого нейрона выходного слоя на его же вход.

6.2. Расстояние или мера Хемминга. Сеть выбирает эталон, для которого расстояние Хэмминга от предъявленного входного вектора путем активации только одного выхода сети (нейрона выходного слоя), соответствующего этому эталону. Расстояние Хемминга равно числу несовпадающих компонент (битов) двух бинарных векторов. Иначе говоря, при использовании двоичных значений (0,1) расстояние Хемминга между двумя векторами

$$\mathbf{y} = (y_1, y_2, \ldots, y_n) \text{ и } \mathbf{x} = (x_1, x_2, \ldots, x_n)$$

определяется в виде:

$$d_H(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} [x_i(1 - y_i) + (1 - x_i) y_i]. \qquad (6.1)$$

При биполярных значениях (± 1) элементов обоих векторов расстояние Хемминга рассчитывается по формуле

$$d_H(\mathbf{y}, \mathbf{x}) = \frac{1}{2}[n - \sum_{j=1}^{n} x_j y_j]. \qquad (6.2)$$

Мера Хемминга равна нулю только когда $\mathbf{y} = \mathbf{x}$. В противном случае она равна количеству битов, на которое различаются два вектора **x** и **y** Соотношение (6.2) доказывается легко непосредственно, а в качестве упражнения полезно проделать переход от выражения (6.2) к формуле (6.1).

6.3. Обучение сети. Фактически первый слой в сети Хемминга отсутствует и она двухслойная. На стадии инициализации сети Хемминга

весовым коэффициентам первого слоя (в двухслойном рассмотрении) присваиваются значения

$$w_{ij}^1 = \frac{x_i^j}{2},\ \theta_j = \frac{n}{2}, \tag{6.3}$$

i=1,2,...,n; j=1,2,...,k; где x_i^j - значение i-го признака (бита) для j-го образца. Весовые коэффициенты тормозящих обратных связей во втором слое полагают равными некоторой небольшой величине ε: $0<\varepsilon<1/k$. Выход нейрона, связанный с его же входом, имеет в начале вес +1.

6.4. Работа сети. Опишем и проанализируем алгоритм функционирования сети Хемминга:

1. На входы сети подается неизвестный вектор $\mathbf{x} = \{x_i\}$, на основе которого рассчитываются выходы состояния нейронов первого слоя

$$y_j = \sum_{i=1}^{n} w_{ij}^1 x_i + \theta_j,\ j=1,...,k. \tag{6.4}$$

Для того, чтобы узнать, что представляет эта величина с точки зрения расстояния Хемминга, подставим в формулу (6.4) весовые коэффициенты (6.3).

Тогда

$$y_j = \sum_{i=1}^{n} \frac{x_i^j}{2} x_i + \frac{n}{2} = \frac{1}{2}[n + \sum_{i=1}^{n} x_i^j x_i]. \tag{6.5}$$

Сравнивая это выражение с формулой (6.2) для расстояния Хемминга, мы получим

$$y_j = n - d_H(\mathbf{x}^j, \mathbf{x}). \tag{6.6}$$

Если ввести нормированные значения выходных сигналов

$$\hat{y}_j = \frac{y_j}{n} = 1 - \frac{d_H(\mathbf{x}^j, \mathbf{x})}{n}, \tag{6.7}$$

то $\hat{y}_j = 1$, если $\mathbf{x}^j = \mathbf{x}$ и $\hat{y}_j = 0$, если $\mathbf{x}^j = -\mathbf{x}$. В остальных случаях значения $\hat{y}_j$ располагаются в интервале [0,1].

Сигналы $\hat{y}_j$ становятся начальными состояниями второго слоя MAXNET на второй фазе функционирования сети. Задача нейронов этого слоя состоит в определении победителя, то есть нейрона, уровень возбуждения которого наиболее близок к 1. Эта задача в алгоритмическом смысле вполне элементарна, но интересно ее решение с помощью нейронной сети. Нейрон - победитель указывает на вектор образа с минимальным расстоянием Хемминга до входного вектора **x**. Процесс определения победителя - это рекуррентный процесс. Для его выполнения веса обратной связи второго слоя выбираются в виде

$$w_{jj}^2 = 1, w_{jl}^2 = -\varepsilon, 0 < \varepsilon < 1/k, \ j \neq l. \qquad (6.8)$$

Рекуррентная формула, определяющая итерационный процесс, имеет вид

$$\hat{y}_j(t+1) = f(\sum_l w_{jl}^2 \hat{y}_l(t)) = f(\hat{y}_j(t) - \varepsilon \sum_{l \neq j} \hat{y}_l(t)). \qquad (6.9)$$

Функция активации нейрона *f*(*y*) слоя MAXNET задается выражением

$$f(y) = \begin{cases} y, y \geq 0 \\ 0, y < 0 \end{cases}. \qquad (6.10)$$

Уравнение (6.9) показывает, что значения на выходе нейронов будут уменьшаться, пока не достигнут 0 во всех кроме одного. Это достаточно очевидно, если занумеровать $\hat{y}_i(1)$ в порядке убывания. Тогда ясно, что «выживет» один нейрон с *i*=1, поскольку снижение значений у остальных нейронов будет происходить быстрее.

Проблема, связанная с сетью Хемминга проявляется в случае, когда зашумленные образы находятся на одинаковом (в смысле Хемминга) расстоянии от двух или более эталонов. В этом случае выбор сетью Хемминга одного из этих эталонов становится совершенно случайным.

Глава 7. Сети Кохонена

7.1. Самоорганизующиеся топологические карты. Сети Кохонена, или самоорганизующиеся карты (Kohonen maps), предназначены для решения задач автоматической классификации, когда обучающая последовательность образов отсутствует (обучение без учителя). Соответственно невозможно и определение ошибки классификации, на минимизации которой построено обучение с учителем (например, в алгоритме обучения с обратным распространением ошибки).

Сеть Кохонена является двухслойной. Она содержит слой входных нейронов и собственно слой Кохонена. Слой Кохонена может быть одномерным, двумерным и трехмерным. В первом случае нейроны расположены в цепочку; во втором – они образуют двумерную сетку (обычно в форме квадрата или прямоугольника), а в третьем – трехмерную систему. Определение весов нейронов слоя Кохонена основано на использовании алгоритмов автоматической классификации (кластеризации или самообучения).

Пусть имеется сеть Кохонена, содержащая *n* входных нейронов и слой Кохонена из *m* выходных нейронов, расположенных в виде прямоугольника (рис. 7.1)

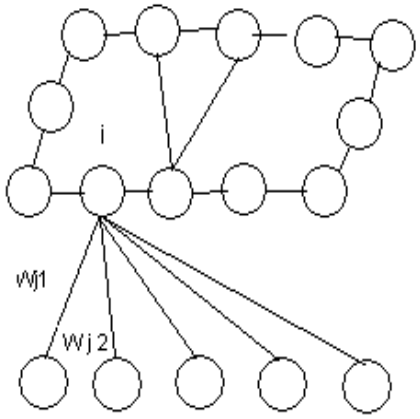


Рис. 7.1. Структура сети Кохонена

7.2. Данные и нейроны. На вход сети подаются последовательно значения векторов $\mathbf{x}_l = (x_1, x_2, \ldots, x_n)_l$, представляющих отдельные

последовательные наборы данных для поиска кластеров, то есть различных классов образов, причем число этих кластеров заранее неизвестно. На рис. 7.1 показаны связи всех входных нейронов лишь с одним нейроном слоя Кохонена. Каждый нейрон слоя Кохонена соединен также с соседними нейронами.

Введем следующие определения. Нейроны входного слоя служат для ввода значений признаков распознаваемых образов. Активные нейроны слоя Кохонена предназначены для формирования областей векторов весовых коэффициентов j-го нейрона слоя Кохенена $\mathbf{w}_j = (w_{j1}, w_{j2}, ..., w_{jn})$, ($j$=1,2,...,$m$), в то время как входной вектор или вектор значений признаков входного образца $\mathbf{x} = (x_1, x_2, ..., x_n)$.

На стадии обучения (точнее самообучения) сети входной вектор $\mathbf{x}$ попарно сравнивается со всеми векторами $\mathbf{w}_j$ всех нейронов сети Кохонена. Вводится некоторая функция близости d (например, в виде евклидова расстояния). Активный нейрон с номером c слоя Кохонена, для которого значение функции близости $d(\mathbf{x}, \mathbf{w}_c)$ между входным вектором $\mathbf{x}$, характеризующим некоторый образ, к векторам $\boldsymbol{\omega}_c$ максимально, объявляется «победителем». При этом образ, характеризующийся вектором $\mathbf{x}$, будет отнесен к классу, который представляется нейроном-«победителем».

7.3. Самообучение сетей Кохонена. Рассмотрим алгоритм самообучения сетей Кохонена. Обозначим функцию близости $z = \|\mathbf{x} - \mathbf{w}\|$. Выигрывает нейрон с

$$\|\mathbf{x} - \mathbf{w}_c\| = \min \|\mathbf{x} - \mathbf{w}_j\|. \qquad (7.1)$$

Для многомерных данных можно (и желательно) использовать нормированные векторы $\mathbf{x}' = \mathbf{x}/\|\mathbf{x}\|$, $\mathbf{w}'_j = \mathbf{w}_j / \|\mathbf{w}_j\|$:

$$x'_i = \frac{x_i}{\sqrt{x_1^2 + x_2^2 + ... + x_n^2}}, \qquad w'_{ij} = \frac{w_{ij}}{\sqrt{\omega_{1j}^2 + \omega_{2j}^2 + ... + \omega_{nj}^2}}. \qquad (7.2)$$

Близость $\mathbf{x}'$ и $\mathbf{w}'$ можно переделить, пользуясь скалярным произведением, как

$$z_j = 1 - (\mathbf{x}', \mathbf{w}'_j) = 1 - \sum_{i=1}^{n} x'_i w_{ij} . \qquad (7.3)$$

На стадии самообучения сети Кохонена осуществляется коррекция весового вектора не только нейрона-«победителя», но и весовых векторов остальных активных нейронов слоя Кохонена, однако, естественно, в значительно меньшей степени – в зависимости от удаления от нейрона-«победителя». При этом форма и величина окрестности вокруг нейрона-«победителя», весовые коэффициенты нейронов которой также корректируются, в процессе обучения изменяются. Сначала начинают с очень большой области – она, в частности, может включать все нейроны слоя Кохонена. Изменение весовых векторов осуществляется по правилу

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \eta(t) d_{cj}(t)[\mathbf{x}(t) - \mathbf{w}_j(t)], \qquad j = 1,2,\ldots, n, \qquad (7.4)$$

где $\mathbf{w}_j(t)$ - значение весового вектора на шаге t самообучения сети, $d_{cj}(t)$ - функция близости между нейронами слоя Кохонена и $\eta(t)$ - изменяемый во времени коэффициент шага коррекции. В качестве $\eta(t)$ обычно выбирается монотонно убывающая функция ($0 < \eta(t) < 1$), то есть алгоритм самообучения начинается сравнительно большими шагами адаптации и заканчивается относительно малыми изменениями.

В результате n-мерное входное пространство R_n отобразится на m-мерную сетку (слой Кохонена). Следует подчеркнуть, что это отображение реализуется в результате рекуррентной (итеративной) процедуры самообучения (unsupervised learning). Отличительная особенность этого отображения – формирование кластеров (clusters) или классов. По завершении процесса самообучения на стадии реального использования сети Кохонена неизвестные входные образы относятся к одному из выявленных кластеров (классов) по близости к некоторому весу, принадлежащему определенному кластеру, выявленному на стадии самообучения.

7.4. Последовательность алгоритма Кохонена.

Шаг 1. Инициализация. Выбираются (случайным образом) начальные значения всех n-мерных весовых векторов слоя Кохонена, а также стартовые значения коэффициента коррекции η и радиуса близости d.

Шаг 2. Выбирается некоторый образ, характеризующийся вектором значений признаков $\mathbf{x}_j$.

Шаг 3. Определяется нейрон-«победитель» с номером c.

Шаг 4. Для нейрона-«победителя» с номером c и для нейронов из радиуса близости вокруг него определяются новые значения весовых коэффициентов (векторов).

Шаг 5. Осуществляется модификация коэффициента коррекции и радиуса близости d.

Шаг 6. Отбирается критерий сходимости, согласно которому происходит «останов» или переход к шагу 2.

Функция близости можно выбрать в виде

$$d_{Gauss}(z) = e^{-z^2},$$

$$d_{mexicanhat}(z) = (1 - z^2)e^{-z^2}$$

$$d_{\cos}(z) = \begin{cases} \cos\left(z\frac{\pi}{2}\right), z \le 1 \\ 0, z > 1 \end{cases}$$

где z – расстояние между нейронами в обычном пространстве, например для двумерного пространства векторов

$$z = \sqrt{(k_{i1} - k_{j1})^2 + (k_{i2} - k_{j2})^2},$$

где k_{i1} и k_{i2} – координаты по оси x и оси y нейрона i, k_{j1} и k_{j2} – аналогично для нейрона j.

Глава 8. Нечеткие множества и лингвистические переменные

8.1. Нечеткие множества. Человеческое мышление характерно тем, что оно позволяет принимать правильные решения в условиях неполной и нечеткой информации. Построение моделей приближенных рассуждений человека, и их использование в интеллектуальных компьютерных системах представляет собой одно из перспективных направлений в области развития интеллектуальных систем.

Основы данного направления заложены L. Zadeh в 1965 г. работой «Нечеткие множества» («Fuzzy sets»). В ней классическое понятие множества расширено путем введения характеристической функции (функции принадлежности элемента множеству), которая может принимать любые значения в интервале [0, 1], а не только значения 0 или 1, то есть значения false и true, как это имеет место для обычных четких множеств. Математическая теория нечетких множеств позволяет описывать нечеткие понятия и значения, а также оперировать этими понятиями и делать нечеткие выводы. Нечеткая логика ближе к человеческому мышлению и языку чем традиционные логические системы.

Пусть U – универсальное множество, x – элемент U, а P – некоторое свойство элемента $P(x)$. Обычное четкое подмножество A универсального множества U состоит из элементов, удовлетворяющих свойству $P(x)$, то есть определяется как упорядоченное множество пар $A=\{\mu_A(x)/x\}$, где $\mu_A(x)$ – характеристическая функция принимающая значение 1, если выполняется условие $P(x)$ и 0, если не выполняется:

$$\mu_A(x) = \begin{cases} 0, \text{ если } \overline{P(x)}, \\ 1, \text{ если } P(x). \end{cases} \qquad (8.1)$$

Нечеткое множество отличается от обычного тем, что для него нет столь определенного ответа, относительно свойства P. Поэтому нечеткое подмножество A универсального множества U определяется как множество упорядоченных пар $A=\{\mu_A(x)/x\}$, где $\mu_A(x)$ – характеристическая функция принадлежности, такая что $\mu_A(x) \in [0, 1]$.

Пример записи нечеткого множества:

Пусть универсальное множество состоит из элементов

$U=\{x_1, x_2, x_3, x_4, x_5\}$, $\mu_A(x_1) = 0{,}1$; $\mu_A(x_2) = 0$; $\mu_A(x_3) = 1$; $\mu_A(x_4) = 0{,}5$; $\mu_A(x_5) = 0{,}9$. Тогда множество A с такой характеристической функцией можно представить в виде: $A=\{0{,}1/x_1; 0/x_2; 1/x_3; 0{,}5/x_4; 0{,}9/x_5\}=0{,}1/x_1 +0/x_2 +1/x_3 +0{,}5/x_4 +0{,}9/x_5$, где «+» не является обозначением операции сложения, а означает объединение элементов. Для непрерывных множеств используется интегральная форма записи $A=\int \frac{\mu_A(x)}{x}dx$, которая также символически задает пары элемент-значение функции принадлежности.

8.2. Основные характеристики нечетких множеств. Величина «супримум» (sup) – точная верхняя граница, sup $\mu_A(x)$, $x\in U$. Если множество конечное, то есть состоит из конечного числа эле ментов, то sup и max – одно и то же. Есть понятие inf (инфимум) – точная нижняя граница множеств. $h(A)=\sup\ \mu_A(x)$ называется высотой нечеткого множества, где $x\in U$. Множество A нормально, если высота множества $h(A)=1$, и множество субнормально, если $h(A) <1$, где $x\in U$. Нечеткое множество A является пустым, если для любого $x\in U$, $\mu_A(x)=0$.

Непустое субнормальное множество можно параметризовать по формуле:

$$\mu_A(x)=\mu_A(x)/h(A), \qquad (8.2)$$

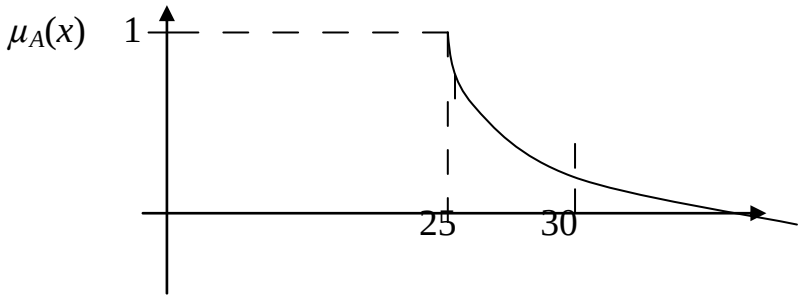
где $x\in U$ и таким образом сделать его нормальным.

Нечеткое множество называется *унимодальным*, если $\mu_A(x)=1$ только на одном $x\in U$. Носителем нечеткого множества A называется его обычное подмножество (четкое) со свойством $\mu_A(x)>0$. Элементы $x\in U$, для которых $\mu_A(x)=0{,}5$ называются точками перехода множества A.

Пример: Пусть универсальное множество U состоит из элементов

$U = \{1, 2,\ldots, 100\}$ и соответствует понятию возраст. Тогда понятие молодой можно определить с помощью следующей функции принадлежности:

$$\mu_{\text{молодой}}(x)= \begin{cases} 1, & x\in[1,25] \\ 1/(1+((x-25)\backslash 5)^2), & x>25. \end{cases} \qquad (8.3)$$



дальше функция убывает

Рис. 8.1. Вид функции принадлежности

Для формирования функции принадлежности используются либо экспертные методы, либо относительные частоты определения принадлежности по данным эксперимента в качестве значений принадлежности.

8.3. Операции над нечеткими множествами. *Включение.* Пусть A и B – нечеткие множества на универсальном множестве U. Тогда A содержится в B, если для любого $\forall x\in U$, $\mu_A(x)\le\mu_B(x)$, $A\subset B$. Говорят также, что B доминирует A.

Равенство. A=B, если $\forall x\in U$, $\mu_A(x)=\mu_B(x)$

Дополнение. A и B дополняют друг друга, если $\forall x\in U$, $\mu_A(x)=1-\mu_B(x)$

$B=\overline{A}$, $A=\overline{B}$.

Очевидно, что ($\overline{\overline{A}}$)=A ($^{=}$ дополнение к дополнению).

Объединение - это наименьшее нечеткое подмножество, включающее как A, так и B с функцией принадлежности:

$A\cup B$, $\mu_{A\cup B}(x)=\max(\mu_A(x),\mu_B(x))$

Пересечение $A \cap B$ - это наибольшее нечеткое множество, содержащее одновременно элементы *A* и *B* с функцией принадлежности:

$\mu_{A\cap B}(x) = \min(\mu_A(x), \mu_B(x))$.

Разность: $A \setminus B = A \cap \overline{B}$ - это подмножество с функцией принадлежности $\mu_{A-B}(x) = \min\{\mu_A(x), 1 - \mu_B(x)\}$.

8.4. Логические операции над нечеткими множествами. Для нечетких множеств можно строить визуальное представление. Рассмотрим прямоугольную систему координат, в которой по оси ординат отложены значения $\mu_A(x)$, а по оси абсцисс откладываются элементы *x*, упорядоченные каким-либо образом. Такое представление делает наглядными простые логические операции над нечеткими множествами. Эти диаграммы являются обобщением диаграмм Венна для четких множеств.

Нечеткие и лингвистические переменные. При описании объектов и явлений с помощью нечетких множеств используются понятия нечеткой и лингвистической переменной.

Нечеткая переменная характеризуется тремя параметрами: <*α*,*U*,*A*>, где

α – наименование переменной;

U – универсальное множество, являющееся областью определения *α*;

A – нечеткое множество на *U*, описывающее ограничения, то есть значения $\mu_A(x)$, на значение нечеткой переменной *α*.

Например:

$$\alpha \in \{\underset{0,2}{Вася}, \underset{0,8}{Петя}, \underset{0}{Маша}, \underset{0,5}{Саша}\} = U$$

Вводится нечеткое множество *A*, $\mu_A(x)$

$$\alpha = 0,2/x_1 + 0,8/x_2 + 0/x_3 + 0,5/x_4$$

Лингвистическая переменная характеризуется набором параметров: <*β*, *T*, *X*, *G*, *M*>, где *β* – наименование лингвистической переменной (возраст); *T* – множество ее значений (терм – множество: молодой, старый), представляющих наименование нечетких переменных, областью определения каждой из которых является множество *X*; *G* – синтаксическая

процедура, позволяющая оперировать элементами терм-множества T (например, образовать понятия «не молодой», «не старый»), в частности, генерировать новые терм-значения. $T \cup G(T)$, где $G(T)$ - множество сгенерированных термов, называется расширенным терм-множеством лингвистической переменной; M – семантическая процедура, позволяющая превратить каждое новое значение лингвистической переменной, образуемое процедурой G в нечеткую переменную, то есть сформировать соответствующее нечеткое множество.

Нечеткая логика. Также как в основе четких множеств лежит четкая логика, в нечетких множествах проявляется нечеткая логика как основа операций над нечеткими множествами. Рассмотрим расширения четких логических операций «не», «и», «или» до нечетких операций. Они соответствуют описанным выше операциям дополнения, пересечения и объединения нечетких множеств. Эти расширения называются нечетким отрицанием, *t*-нормой и *S*-нормой соответственно.

В нечетком пространстве логических значений число состояний неограниченно велико, поэтому невозможно описать нечеткие логические операции с помощью таблицы истинности, как в случае двузначной логики.

8.5. Нечеткие правила вывода. Правило вывода:

Правило «е*сли x это A*, то *y это B*» называется нечеткой импликацией $A \rightarrow B$. A и B - лингвистические значения (значения лингвистической переменной), определенные нечетким способом через функции принадлежности для переменных. Часть импликации «*x это A»*, называется условием , а «*y это B»*, называется следствием (заключением). Обобщение для *n*-мерного вектора: $x=\{x_1,x_2,\ldots,x_n\}$, «если x_1 *это* A_1, x_2 *это* A_2, … , x_n *это* A_n», то «*y это B*».

Каждое из условий «x_i *это* A_i» задается с помощью коэффициентов принадлежности $\mu_{A_i}(x_i)$, i = 1, 2,…, n. «*y это B*» описывается функцией принадлежности $\mu_B(y)$. Совокупность *n* условий можно интерпретировать в

форме логического произведения соответствующей операции «и». В этом случае

$$\mu_A(x) = \min\{\mu_{A_i}(x_i)\},\ i = 1,2,\text{K}\ ,n\,. \qquad (8.4)$$

Условия «и» можно оценивать также в форме алгебраического произведения

$$\mu_A(x) = \prod \mu_{A_i}(x_i),\ \ i = 1,2,\text{K}\ ,n. \qquad (8.5)$$

Каждой импликации $A \rightarrow B$ можно приписать значения функции принадлежности $\mu_{A\rightarrow B}(x,y)$ в зависимости от способа задания. В схеме Мамдани-Заде

$$\mu_{A\rightarrow B}(x,y) = \min\{\mu_A(x), \mu_B(y)\}. \qquad (8.6)$$

Для варианта задания правила вывода в форме алгебраического произведения

$$\mu_{A\rightarrow B}(x,y) = \mu_A(x)\cdot\mu_B(y)\,. \qquad (8.7)$$

8.6. Система нечеткого вывода. Элементы теории нечетких множеств, правила импликации и нечетких рассуждений образуют систему нечеткого вывода.

Система нечеткого вывода содержит:

1. Множество используемых нечетких правил, которым соответствуют операции нечетких множеств.
2. Базу данных, содержащую описание функций принадлежности.
3. Механизмы вывода и агрегирования, которые формируется применяемыми правилами импликации.

При технической реализации систем нечеткой логики в качестве входных и выходных сигналов выступают измеряемые величины, однозначно сопоставляющие входным значениям соответствующие выходные величины.

Для обеспечения взаимодействия четких и нечетких систем вводится преобразование четких величин к нечетким величинам, называемое *фаззификатором*. Оно преобразует точное множество входных данных в

нечеткое множество, определяющееся с помощью функции принадлежности. Обратную задачу решает *дефаззификатор* - преобразователь нечетких множеств в конкретное значение выходной переменной.

Фаззификация. Пусть данные представляются в виде *n*-мерного вектора $x=\{x_1, x_2,\ldots,x_n\}$ и $x \in A$.

Функцию принадлежности можно построить различными способами:

1. Часто используются *функции Гаусса*

$$\mu_A(x) = \exp\left(-(x-c)^2 / \sigma^2\right), \tag{8.8}$$

где c – центр нечеткого множества, σ - коэффициент ширины.

2. С помощью *симметричной треугольной функции*

$$\mu_{\grave{A}}(x) = \begin{cases} 1-|x-d|/d, x \in [c-d, c+d] \\ 0, \quad \ddot{a}\ddot{e}\ddot{y} \quad \ddot{a}\eth\acute{o}\tilde{a}\grave{e}\tilde{o} \quad x \end{cases}. \tag{8.9}$$

Дефаззификация обеспечивает трансформацию нечеткого множества с функцией принадлежность $\mu(y)$ в точное значение.

Имеются различные способы осуществления такого преобразования:

1. Деффазификация относительно центра области, путем вычисления интеграла для непрерывных множеств:

$$y_C = \frac{\int \mu_C(y) \cdot y dy}{\int \mu_C(y) dy}. \tag{8.10}$$

Для дискретных множеств:

$$y_C = \frac{\sum_{i=1}^{k} \mu_C(y_i) y_i}{\sum_{i=1}^{k} \mu_C(y_i)}, \; y_i : i = 1,2,\ldots,k. \tag{8.11}$$

2. Деффазификации относительно среднего максимума:

$$y_m = \frac{\sum_{i=1}^{m} y_i}{m}, \tag{8.12}$$

где m – количество точек, в которых функция $\mu(y_i)$ достигает локально-максимальных значений. Возможны и другие непротиворечивые способы определения точного значения.

8.7. Модель Мандани-Заде – универсальный аппроксиматор. Модели нечеткого вывода позволяют описать выходной сигнал многомерного процесса как нелинейную функцию входных переменных x_i, где i=1,2,….n и параметров нечеткой системы.

В модель Мандани – Заде вводится конечное число M правил, где каждое из правил, определяется уровнем активации из условия:

$$\mu(y_i) = \prod_{j=1}^{M} \mu_{A_i}(x_j), \tag{8.13}$$

где y_i - значение величины y, при котором $\mu(y_i) = \max$.

Пусть c_i - центр нечетких множества, соответствующего заключению для i-го правила вывода. Тогда дефаззификация относительно среднего центра дает:

$$y = \frac{\sum_{i=1}^{M} c_i \mu(y_i)}{\sum_{i=1}^{M} \mu(y_i)}. \tag{8.14}$$

Модель Мандани – Заде близка к структуре классических нейронных сетей с радиальными базисными функциями. Сети, построенные на основе подобных нечетких моделей, называются нечеткими нейронными сетями.

8.8. Нечеткие сети TSK (Такаши-Сугено-Канга). Схема вывода для модели TSK при использовании M правил и n переменных x_j использует в качестве выхода четкую функцию, вычисляемую с помощью весов полученных, на основе нечетких правил:

$$\textit{если } (x_1 \textit{ это } A_1) \wedge (x_2 \textit{ это } A_2) \wedge \ldots \wedge (x_n \textit{ это } A_n) \textit{ тогда}$$

$$y = p_0 + \sum_{j=1}^{n} p_j x_j. \tag{8.15}$$

Коэффициенты p_0, p_1, K, p_n подбираются в результате обучения сети. Условие «x_j *это A*» реализуется функцией фаззификации:

$$\mu_A(x_j) = \frac{1}{1 + \left(\frac{x_j - c_j}{\sigma_j} \right)^2}. \tag{8.16}$$

При наличии *M* правил вывода, после агрегирования выходной результат сети имеет вид:

$$y(x) = \frac{\sum_{i=1}^{M} w_i y_i(x)}{\sum_{i=1}^{M} w_i}, \tag{8.17}$$

$$y_i(x) = p_{i0} + \sum_{j=1}^{n} p_{ij} x_j . \tag{8.18}$$

Веса w_i интерпретируются как значимость компонентов $\mu(y_i)$.

Формулам (8.17), (8.18) можно сопоставить многослойную нейронную сеть TSK:

1. Первый слой выполняет фаззификацию каждой переменной и определяет параметры c_j и σ_j, подлежащие адаптации в процессе обучения.

2. Второй слой определяет агрегирование определенных переменных, вычисляя результирующие значения коэффициента принадлежности:

$$w_j = \mu(y_j). \tag{8.19}$$

3. Третий слой рассчитывает значения y_i по формуле:

$$y_i(x) = p_{i0} + \sum_{j=1}^{n} p_{ij} x_i . \tag{8.20}$$

4. Четвертый слой составляют два нейрона сумматора, первый из которых вычисляет числитель, а второй знаменатель формулы

$$y(x) = \frac{\sum_{i=1}^{M} w_i y_i(x)}{\sum_{i=1}^{M} w_i}. \tag{8.21}$$

5. Пятый слой заканчивает вычисление $y(x)$ делением.

Обучение нечетких сетей осуществляется с помощью гибридных алгоритмов. Они включают в себя, обычно, решение систем линейных уравнений и метод наискорейшего спуска. Нечеткие нейронные сети используются для построения мягких экспертных систем и часто применяются в сочетании с генетическими алгоритмами.

9. Экспертная оценка конкурентоспособности товаров на основе нечеткого вывода

9.1. Конкурентоспособность товара. Конкурентоспособность товара в общем случае определяется тремя необходимыми элементами [1,2]: свойствами данного товара, свойствами конкурирующих товаров, особенностями потребителей. Суть конкуренции выражается появлением новых конкурентов, появлением товаров или услуг-заменителей, способностью поставщиков комплектующих изделий торговаться, способностью покупателей торговаться, соперничеством уже имеющихся конкурентов. Конкурентоспособный товар обладает конкурентными преимуществами, которые делятся на два основных вида: более низкие издержки и дифференциация товаров [3,4]. Низкие издержки отражают способность фирмы разрабатывать, выпускать и продавать сравнимый товар с меньшими затратами, чем конкуренты. Дифференциация – это способность обеспечить покупателя товаром нового качества, особых потребительских свойств или с послепродажным обслуживанием. Следует отметить, что чаще всего рассматриваются только свойства данного товара и свойства конкурирующих товаров. Расчетные способы определения конкурентоспособности товара оперируют именно этими группами показателей – параметров качества (технических) и экономических параметров.

От выбора базы сравнения в значительной степени зависит правильность результата оценки конкурентоспособности и принимаемые в

дальнейшем решения. Базой сравнения могут выступать: потребность покупателей; величина полезного эффекта; конкурирующий товар; гипотетический образец; группа аналогов. В случае, когда базой сравнения является потребность покупателей, осуществляется выбор номенклатуры и установление величин параметров потребности покупателей в оцениваемой и конкурирующей продукции, которыми потребитель пользуется при оценке продукции на рынке, а также весомости этих параметров в общем их наборе. Когда за базу сравнения принимается величина полезного эффекта продукции, а также возможные расходы потребителя на ее приобретение, выделяются полезный эффект в качестве эталона или сумма соответствующих средств. Если продукция имеет конкурента, то товар-образец моделирует потребность и выступает в качестве требований, которым должна удовлетворять продукция, подлежащая оценке.

В качестве базы сравнения может выступать гипотетический образец, который представляет собой среднее значение параметров группы изделий. Такая процедура используется, когда информации по конкретному образцу-аналогу недостаточно. Эта оценка должна рассматриваться как ориентировочная и подлежащая дальнейшему уточнению. Чаще за базу сравнения принимается группа аналогов с согласованными классификационными параметрами, из которой выбираются наиболее представительные, а затем имеющие наилучшую перспективу для дальнейшего расширения объема продаж. Целью данной работы является разработка системы оценки качества товара, основанной на экспертном подходе.

9.2. Критерии оценки конкурентоспособности товаров. Стандартно оценка конкурентоспособности товара производится путем сопоставления параметров анализируемой продукции с параметрами базы сравнения. Сравнение проводится по группам технических и экономических параметров. При оценке используются дифференциальный и комплексный методы оценки. Дифференциальный метод оценки конкурентоспособности, основан

на использовании единичных параметров анализируемой продукции и базы сравнения. Если за базу оценки принимается потребность, то расчет единичного показателя конкурентоспособности производится по формуле:

$$q_i = P_i / P_{i0}, \qquad (9.1)$$

где q_i – единичный показатель конкурентоспособности по i-му параметру (i = 1, 2, 3, ..., n); P_i – величина i-го параметра для анализируемой продукции; P_{i0} – величина i-го параметра, при котором потребность удовлетворяется полностью; n – количество параметров.

Дифференциальный метод позволяет констатировать факт конкурентоспособности анализируемой продукции или наличия у нее недостатков по сравнению с товаром-аналогом. Он не учитывает влияние на предпочтение потребителя при выборе товара весомости каждого параметра. Для устранения этого недостатка используется комплексный метод оценки конкурентоспособности с расчетом группового показателя. Расчет группового показателя по нормативным параметрам производится по формуле:

$$I = \prod_i q_i , \qquad (9.2)$$

где I – групповой показатель конкурентоспособности. Отличительной особенностью данной формулы является то, что если хотя бы один из единичных показателей равен 0, что означает несоответствие параметра обязательной норме, то групповой показатель также равен 0. Очевидно, что товар при этом будет неконкурентоспособен.

Расчет группового показателя по техническим параметрам (кроме нормативных) производится по формуле:

$$I_t = \sum_i q_i a_i , \qquad (9.3)$$

где I_t – групповой показатель конкурентоспособности по техническим параметрам; a_i – весомость i-го параметра в общем наборе из n технических параметров, характеризующих потребность. Полученный групповой

показатель I_t характеризует степень соответствия данного товара существующей потребности по всему набору технических параметров, чем он выше, тем в целом полнее удовлетворяются запросы потребителей. Основой для определения весомости каждого технического параметра в общем наборе являются экспертные оценки, основанные на результатах маркетинговых исследований. Иногда в целях упрощения расчетов и проведения ориентировочных оценок из технических параметров может быть выбрана наиболее весомая группа или применен комплексный параметр – полезный эффект, который в дальнейшем участвует в сравнении.

Расчет группового показателя по экономическим параметрам производится на основе определения полных затрат потребителя на приобретение и эксплуатацию продукции. Полные затраты потребителя определяются по формуле:

$$Z = Z_0 + \sum_{i=1}^{T} C_i \,, \qquad (9.4)$$

где Z – полные затраты потребителя на приобретение и эксплуатацию продукции, Z_0 – единовременные затраты на приобретение продукции, C_i – средние суммарные затраты на эксплуатацию продукции, относящиеся к *i*-му году ее службы, T – срок службы, i – год эксплуатации. При этом

$$C_i = \sum_{j=1}^{n} C_{ij} \,, \qquad (9.5)$$

где C_{ij} – эксплуатационные затраты по *j*-ой статье; n – количество видов эксплуатационных затрат.

Расчет группового показателя по экономическим параметрам производится по формуле:

$$I_e = \frac{\mathrm{Z}}{\mathrm{Z_c}}, \qquad (9.6)$$

где I_e – групповой показатель по экономическим параметрам, Z, Z_c – полные затраты потребителя соответственно по оцениваемой продукции и

образцу. В случае необходимости учитывается коэффициент приведения эксплуатационных затрат.

Расчет интегрального показателя конкурентоспособности производится по формуле:

$$K = I \cdot I_t / I_e, \qquad (9.7)$$

где K – интегральный показатель конкурентоспособности анализируемой продукции по отношению к изделию-образцу. По смыслу показатель K отражает различие между сравниваемой продукцией в потребительском эффекте, приходящемся на единицу затрат покупателя по приобретению и потреблению изделия. Если $K<1$, то рассматриваемый товар уступает образцу по конкурентоспособности, а если $K>1$, то превосходит. При равной конкурентоспособности $K=1$. Если анализ проводится по нескольким образцам, интегральный показатель конкурентоспособности продукции по выбранной группе аналогов может быть рассчитан как сумма средневзвешенных показателей по каждому отдельному образцу.

Смешанный метод оценки представляет собой сочетание дифференциального и комплексного методов. При смешанном методе оценки конкурентоспособности используется часть параметров рассчитанных дифференциальным методом и часть параметров рассчитанных комплексным методом. Данный подход является общеупотребительным, но следует заметить его существенный недостаток – потребительские свойства товара и их набор определяются без учета мнения потребителя. Как следует из приведенных расчетных соотношений, предполагается, что улучшение любой из характеристик товара автоматически повышает его конкурентоспособность. Так, например, если бетонный блок окажется на несколько килограммов легче базового образца, то это, в соответствии с приведенным подходом, означает повышение конкурентоспособности товара. На самом деле это не так однозначно. Возможно, потребитель использует такие блоки в качестве противовесов, обеспечивающих устойчивость. Иначе говоря, улучшение в определенном смысле

характеристик товара по сравнению с базовым образцом вовсе не гарантирует появление конкурентных преимуществ. Решающая роль в оценке преимуществ или недостатков товара остается за потребителем. Совокупность качественных и стоимостных характеристик товара, способствующих созданию превосходства данного товара перед товарами-конкурентами в удовлетворении конкретной потребности покупателя, определяет конкурентоспособность товара.

Экономическая практика показала, что потребители на рынке не выступают единым целым – они по-разному реагируют даже на один и тот же товар с одними и теми же свойствами. Это обстоятельство учитывается маркетологами при сегментировании рынка и позиционировании товара. Таким образом, чтобы определить конкурентоспособность товара, нужно сравнить его свойства со свойствами конкурентов, а также изучить поведение потребителей и их реакцию на товар.

9.3. Основные характеристики системы нечеткого вывода. Проблема конкуренции и конкурентоспособности столь разнообразна, что нам необходимо более четко сформулировать предмет и цель дальнейшего исследования. Мы ограничимся только микроэкономическим уровнем, сосредоточив внимание на конкуренции товаров. В определенной мере это ограничивает получаемые результаты и выводы, поскольку роль региональных и мировых факторов, а также влияние кооперации на процессы конкуренции весьма велики. Мы будем в дальнейшем предполагать, что эти факторы являются заданными и играют роль фона, на котором разворачивается микроэкономическая конкуренция товаров.

Целью работы является придание уже существующим моделям формы, пригодной не только для качественной характеристики конкурентной ситуации, но ее количественного описания, имея в виду проблему оценки конкурентоспособности товара. Ранее мы привели методику численного расчета соответствующего показателя и дали общую характеристику конкурентоспособности. Однако недостатком такой методики является ее

жесткость, в силу чего ее трудно модифицировать для более тонкого учета деталей конкретной ситуации, включая чисто качественные факторы, такие как общая структура платежеспособности населения в данном районе. Стандартные математические средства не позволяют учитывать одновременно как количественные, так и качественные характеристики, при том, что конечный результат необходимо представить в понятной и универсальной числовой форме, пригодной для сравнения одного товара с другим. Подобную возможность предоставляет аппарат нечетких множеств и нечеткого вывода [6-11], которым мы и воспользуемся для решения данной задачи.

Заметим сразу, что конкурентоспособность товара еще не означает конкурентоспособность производящей его фирмы, более того этот фактор может означать даже большой объем убыточных продаж и вести в конечном итоге к разорению. Мы пока будем рассматривать вопрос изолированно. Описанный выше аналитический способ оценки конкурентоспособности, несмотря на его логичность и обоснованность обладает определенной умозрительностью. Наиболее естественным критерием конкурентоспособности может быть легко измеряемый параметр, относительно которого существует понятная и непротиворечивая методика измерения. При выборе товара индивидуальный покупатель руководствуется известным отношением цена/качество, стремясь его уменьшить. Однако даже этот простой критерий приводит к затруднениям при оценке конкурентоспособности товара. Во-первых, непонятно в чем измерять качество товара и в каких единицах. Во-вторых, если это отношение зафиксировать, то можно получить нелепый вывод о том, что дешевый товар плохого качества будет так же покупаться, как и товар с умеренной ценой и нормального качества. Поэтому и такой критерий страдает субъективностью, к тому же непонятно, как его можно непосредственно измерить.

Конкурентоспособность товара, в конечном счете, определяется выбором покупателя и может резко меняться под действием изменения

спроса, обусловленного косвенными факторами. Например, резкое снижение тарифов на перевозки может повышать конкурентоспособность цементных смесей из относительно удаленных районов с хорошей сырьевой и энергетической базой. Что же мы можем взять в качестве параметра оценки конкурентоспособности товара в линейке однородных товаров? Соответствующим критерием может служить, на наш взгляд, только рыночный показатель – относительная частота продаж, которую легко измерять по статистическим данным. Соответствующий эквивалентный параметр оценки можно определить как вероятность выбора потребителем данного товара, или лучше как степень принадлежности товара к конкурентной продукции, оценкой которой мы и займемся далее. Для решения данной задачи мы создадим довольно компактную экспертную систему, действующую на основе нечеткой логики.

9.4. Структура экспертной системы. Проведем оценку товара по пяти параметрам:

1) цена товара;
2) потребительские качества товара;
3) эстетика и реклама товара;
4) новизна товара;
5) категория покупателей.

Будем описывать каждый из этих параметров с помощью нечетких множеств с диапазонами значений аргумента 1-10.

1) Цена товара будет оцениваться тремя значениями лингвистической переменной {низкая (low), средняя (middle), высокая (high)}.
2) Качество товара будет оцениваться четырьмя значениями лингвистической переменной {не очень хорошее (poor), хорошее (good), первоклассное (first class), высшее (perfect)}.
3) Эстетические качества товара и его имидж будет оцениваться тремя значениями лингвистической переменной {не эстетичный (non esthetics), эстетичный (esthetics), отличное (excellent}.

4) Новизна товара будет оцениваться тремя значениями лингвистической переменной {старый (old), новый (new), инновационный (innovative)}.
5) Категория покупателей будет оцениваться значениями лингвистической переменной {бедный (poor), средний класс (middle class), богатый (rich)}.

Результат будем описывать лингвистической переменной «предпочтение» {редко (seldom), средне (middle), часто (often)}и диапазоном результата 100.

Общее количество разных случаев, обрабатываемых экспертной системой, составит $3\times4\times3\times3\times3=324$. Обучение системы такого размера представляет существенные трудности. Поэтому естественно искать возможности упрощения системы. Такую возможность дает метод фокусирования на сегментах потребителей с разной покупательной способностью. Разделяя потребителей на три категории, можно получить три независимые системы, каждая из которых содержит уже 108 вариантов. В этом случае настройка каждой из подсистем по отдельности и их функционирование становятся существенно проще и значительно ускоряются. Дальнейшее существенное снижение размерности системы, на наш взгляд, приведет к неоправданному упрощению отображения и потери адекватности описания. Таким образом, мы остановились на создании трех параллельных экспертных систем для категорий покупателей:

- бедный (poor),
- средний класс (middle class),
- богатый (rich).

Для категории качества мы оставим два значения: не эстетичный (non esthetics) и эстетичный (esthetics). Такое усечение вариантов мы связываем с агрегированием переменных эстетичный (esthetics) и отличное (excellent} в одну переменную – эстетичный (esthetics). Цена товара будет оцениваться нами тремя значениями лингвистической переменной {низкая (low), средняя (middle), высокая (high)}.

Результаты проведенного анализа мы представим в виде итоговой таблицы.

Таблица: Входные и выходные лингвистические переменные оценки конкурентоспособности товара на основе нечеткой логики.

Название переменной	Значение переменной		
Цена товара	низкая (low)	средняя (middle)	высокая (high)
Качество товара	не очень хорошее (poor)	хорошее (good)	первоклассное (first class)
Эстетика товара	не эстетичный (non esthetics)	Эстетичный (esthetics)	
Новизна товара	старый (old)	новый (new)	инновационный (innovative)
Предпочтение (выходная переменная)	**редко (seldom)**	**средне (middle)**	**часто (often)**

Комбинаторика входных параметров дает $3\times3\times2\times3=54$ варианта, что вполне приемлемо для осуществления экспертной настройки. Для подобных систем, которые настраиваются экспертным путем, а не в результате самообучения предпочтительно использовать схему вывода по правилам Мамдани-Заде.

9.5. Результат построения экспертной системы. В качестве примера ниже приведена экспертная система, оценивающая конкурентоспособность товара, при покупке представителями среднего класса на основе треугольных функций принадлежности. Экспертная система реализована средствами MATLAB. На рис. 9.1 показана принципиальная схема экспертной системы, состоящая из четырех входных блоков параметров, блока нечеткого вывод Мамдани-Заде, в котором формируется искомая оценка, и выходного блока оценки предпочтений. Выбор вида функций принадлежности и настройка

правил вывода осуществлялась экспертным способом. Результат и вид сформированной системы показаны на рис. 9.2, которая при конкретных выбранных входных параметрах дает довольно низкую оценку покупательских предпочтений 29,2 по сто бальной шкале оценок. Этот показатель вполне сочетается с ожиданиями оценки для этого случая (низкая цена при низком качестве). На рис. 9.3 показана поверхность отклика в наиболее существенных координатах «цена-качество», которая демонстрирует очевидную приверженность данной категории покупателей к довольно качественным товарам по умеренным ценам.

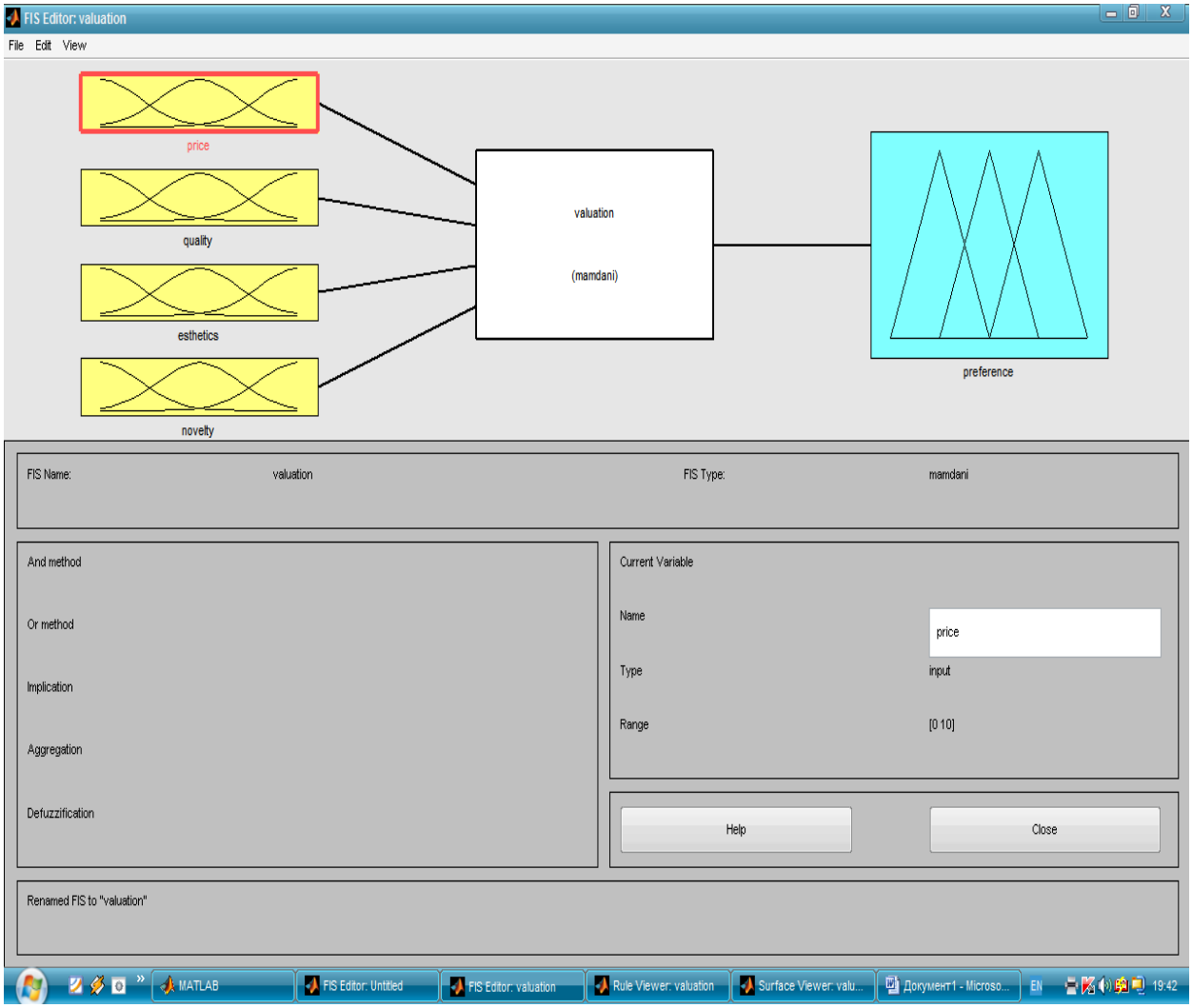


Рис. 9.1. Общая структура экспертной системы

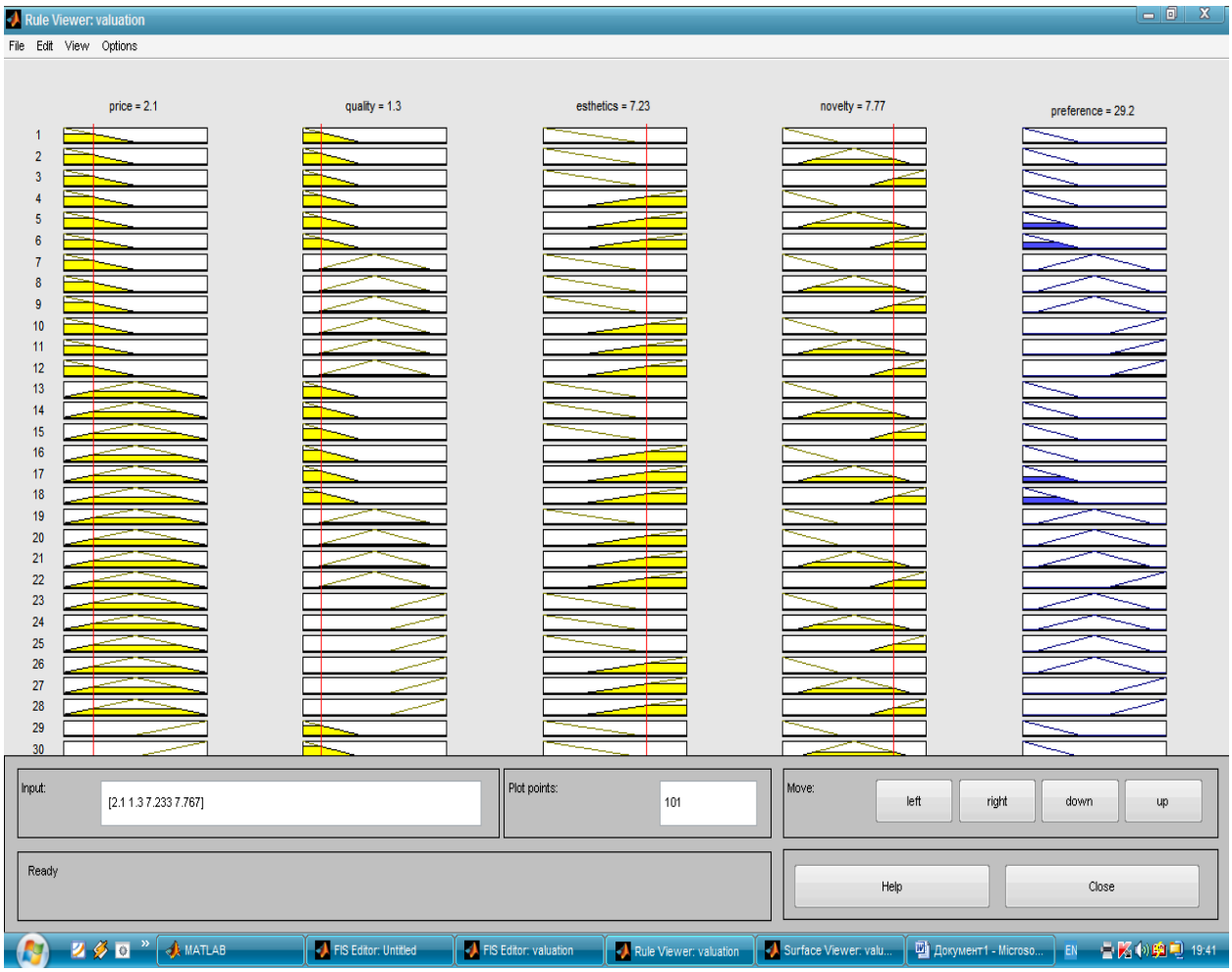


Рис. 9.2. Правила нечеткой логики

Если степень предпочтения для категории покупателей i составляет q_i, а распределение по категориям покупателей p_i, то результирующая оценка товара определяется величиной

$$q = \sum_i p_i q_i \,, \qquad (9.8)$$

что является разумной оценкой качества товара с учетом представленности референтных групп.

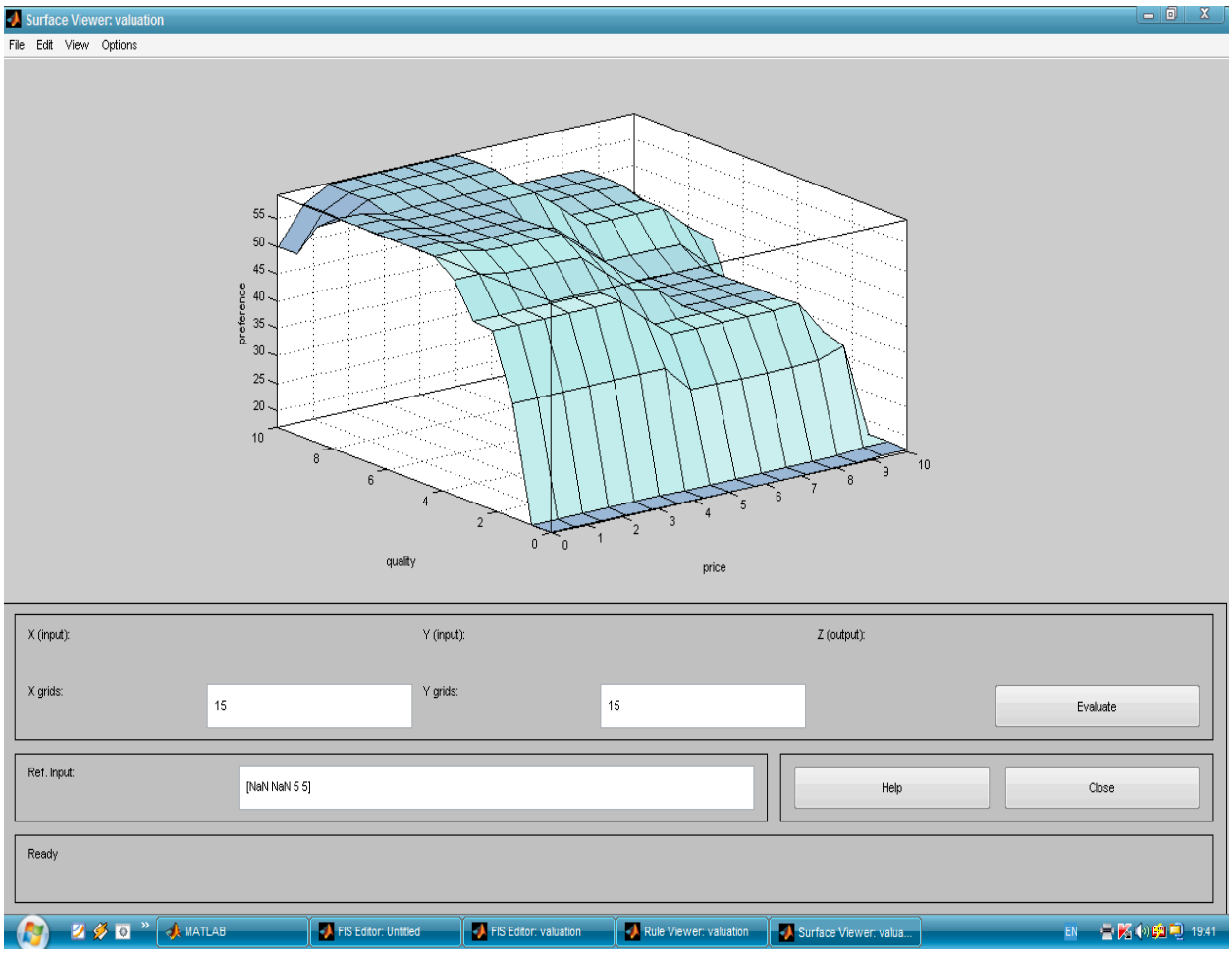


Рис. 9.3. Функция отклика для координат «цена-качества»

Построенную экспертную систему нетрудно модифицировать под другие правила вывода, соответствующие другим экспертным оценкам, а также другим оценочным параметрам.

Результат анализа существующих систем оценок конкурентоспособности товара и их сравнение с опытом построения экспертной системы на основе нечеткой логики и нечеткого вывода показывают, что традиционная методика уступает в гибкости и способности учитывать не только количественные, но и качественные характеристики товара и рынка. Экспертная система, действующая на основе нечеткой логики, имеет наглядную структуру, которая полностью отображается графическим интерфейсом. Система допускает возможность быстрой перенастройки параметров и изменения правил вывода без переписывания программного кода. При создании более совершенной экспертной системы

оценки конкурентоспособности товара целесообразно опираться на опыт и знания маркетологов, специализирующихся по определенной товарной группе. Несмотря на всю важность конкурентоспособности товара для фирмы его производящей, этот параметр является лишь одним в общей проблеме конкурентоспособности фирм, которая требует отдельного рассмотрения, поскольку именно конкуренция фирм наиболее сильно влияет на движение товаров, услуг и финансов в современной рыночной экономике.

ЛИТЕРАТУРА

1. Арнольд В.И. *Теория катастроф*. М: Наука, 1990. – 128 с.
2. Сигорский В.П. *Математический аппарат инженера.* Киев: Техника, 1975. – 768 с.
3. Волков И.К., Загоруйко Е.А. *Исследование операций.* М.: изд-во МГТУ им. Н.Э.Баумана, 2004. – 440 с.
4. Оуэн Г. *Теория игр.* М.: Едиториал УРСС, 2004. – 216 с.
5. Бурков В.Н., Заложнев А.Ю., Новиков Д.А. *Теория графов в управлении организационными системати.* М.: Синтег, 2001. – 124с.
6. Шикин Е.В. *Исследование операций.* М.: Изд-во Проспект, 2006. – 280с.
7. Харари Ф. *Теория графов.* М.: Едиториал УРСС, 2003. – 296 с.
8. Гладков Л.А., Курейчик В.В., Курейчик В.М. *Генетические алгоритмы.* М.: Физматлит, 2006. – 320 с.
9. Кузнецов О.П. *Дискретная математика для инженера.* – М.: Лань, 2005.
10. Чернавский Д.С. *Синергетика и информация* (*динамическая теория информации*). М.: Едиториал УРСС, 2004. – 288 с.
11. Эбелинг В., Энгель А., Файстель Р. *Физика процессов эволюции*. – М.: Едиториал УРСС, 2001. – 328 с.
12. Арбиб М. *Мозг, машина и математика*. – М.: Наука, 1968. – 224 с.
13. Тойнби А.Дж. *Цивилизация перед судом истории*. – М.: Айрис-пресс, 2003. – 592 с.

14. Хантингтон С. *Столкновение цивилизаций*. – М.: АСТ, 2005. – 603 с.
15. Вернадский В.И. *Биосфера и ноосфера*. – М.: Айрис-пресс, 2003. – 576 с.
16. Кун Т. *Структура научных революций*. – М.: АСТ, 2003. – 365 с.
17. Поппер К.Р. *Объективное знание. Эволюционный подход*. – М: Эдиториал УРСС, 2002. – 384 с..
18. Медоуз Д. *Азбука системного мышления*. – М.: Бином, 2011. – 343 с.
19. Занг В.-Б. *Синергетическая экономика*. – М.: Мир, 1999. – 335 с.
20. Евин И.А. *Синергетика мозга.* – Москва- Ижевск: НИЦ «РХД», 2005. – 198 с.
21. Хайкин С. *Нейронные сети.* – М.: Вильямс, 2006. – 1104 с.
22. Кохонен Т. *Самоорганизующиеся карты.* – М.: Бином, 2008. – 655 с.
23. Пегат А. *Нечеткое моделирование и управление.* – М.: Бином, 2009. – 798 с.
24. Медоуз Д. *Азбука системного мышления.* – М.: Бином, 2011. – 343 с.
25. Мюррей *Дж. Математическая биология.* Т.1, Т.2. – М.: НИЦ «РХД». 2009.
26. Вайдлих В. *Социодинамика.* – М.: Либроком, 2010. – 480 с.
27. Суровцев И.С., Клюкин В.И., Пивоваров Р.П. *Нейронные сети.* – Воронеж: ВГУ, 1994. – 224 с.
28. Головинский П.А. *Математические модели.* – М.: Либроком, 2012. Ч. 1. – 240 с; Ч. 2. – 232 с.

Содержание

1. Моделирование
2. Понятие о математической статистике
3. Определение вероятности
4. Условная вероятность
5. Случайные величины
6. Нормальное распределение
7. Распределение χ^2 (Хи-квадрат)
8. Корреляция

Глава 2. Линейный регрессионный анализ

2.1 Приближение табличных значений функций

2.2 Нелинейная регрессия

2.3 Оценка точности регрессии

Глава 3. Временные ряды

3.1 Характеристики временных рядов

3.2 Анализ временных рядов

3.3. Анализ случайной компоненты ряда

3.4. Практический анализ и построение прогноза

Глава 4. Многомерный статистический анализ

4.1 Многомерные данные

4.2 Метрика

4.3 Факторный анализ

4.4 Статистическое распознавание катастроф

Глава 5. Методы исследования операций

5.1 Основные понятия исследования операций

5.2 Задача о составлении рациона

5.3 Задача о быстродействии

5.4 Задача о выборе наилучшей стратегии

5.5 Транспортная задача

5.6 Задача об использовании ресурсов

5.7 Задача составления расписаний

5.8 Постановка задач оптимизации

Глава 6. Линейное программирование

6.1 Постановка задачи

6.2 Геометрическая интерпретация

Глава 7. Сети и графы

7.1. Задачи о сетях

7.2. Общие свойства графов

7.3. Задание графа матрицами

7.5. Пути и связность в графе

7.6. Деревья

7.7. Планарный граф

7.8. Стратегии поиска в пространстве состояний

7.10. Эвристический поиск

Глава 8. Оптимизационные задачи на графах

8.1. Порождающие деревья

8.2. Задача о минимальном порождающем дереве

8.3. Алгоритм построения минимального остова

8.4. Задача о кратчайшем маршруте между выбранными вершинами

8.5. Задача о максимальном потоке

8.6. Реализация сетей в трехмерном пространстве

8.7. Феномен «тесного мира»

Глава 8. Принятие решений при неопределенности целей

9.1. Противоречивость целей

9.2. Линейная свертка

9.3. Использование контрольных показателей

9.4. Простейший способ преодоления неопределенности целей

9.6. Компромиссы Парето

Глава 11. Динамическое программирование

11.1. Принцип оптимальности

11.2. Задача о распределении ресурсов

Глава 12. Элементы теории игр

12.1. Конфликты как игры

12.2. Основное неравенство и игра с седловой точкой

12.3. Игры с вероятностным выбором стратегии.

12.4. Выбор стратегии

Глава 14. Генетические алгоритмы и эволюционное программирование

14.1.Генетические понятия

14.2. Генетический алгоритм

14.3. Эволюция в популяции

14.4. Канонический генетический алгоритм

14.5.Оператор кроссинговера

14.6. Оператор мутации

Заключение

Подход с позиций системного анализа требуется и оказывается чрезвычайно полезным при исследовании самых разнообразных сложных систем. Число таких систем, как в общественных образованиях, так и в технических устройствах чрезвычайно велико, и они отличаются необычайным разнообразием. Все это делает системный анализ труднообозримой дисциплиной с нечетко очерченными границами. В него легко интегрируются самые разнообразные теории и методы. Такая объемность предмета исследований затрудняет и введение единых устоявшихся представлений о системном анализе. Несмотря на это пользу от системного подхода вполне осознали руководители самого разного уровня. Ни один серьезный проект или сценарий развития большой фирмы, страны или мира в целом сегодня немыслим без широкого применения методов системного анализа. Системный анализ повлиял самым существенным образом на проблемы ведения войны, и теперь никакой крупный военный конфликт не начинается развитыми странами без оценок динамики и последствий предполагаемого вооруженного столкновения. Системный анализ самым непосредственным образом повлиял на подписание договоров об ограничении стратегических вооружений и запрет ядерных испытаний в атмосфере, воде и космосе. Связанные с этим исследования стали одной из важнейших ступеней к международному признанию системного анализа как новой науки. Среди создателей системного анализа обязательно нужно отметить российского академика Н.Н. Моисеева и американца Дж. Форрестера из Массачусетского технологического института, впервые предложившего компьютерную модель мировой динамики.

Одним из практических следствий возникновения системного анализа в области мировой политики стало возникновение "Римского Клуба" – международной общественной организации, объединяющей около семидесяти предпринимателей, управляющих, политических деятелей, высокопоставленных служащих, экспертов, деятелей культуры, ученых из

стран Западной Европы, Северной и Южной Америки, Японии. Среди целей "Римского Клуба" – поиск методик анализа глобальных проблем человечества, связанных с ограниченностью ресурсов Земли, ростом производства и потребления, являющимися принципиальными пределами роста.

Вызовы, с которыми сталкивается Россия, непременно потребуют все более широкого использования системного подхода. Одним из российских лидеров системного анализа в России является заместитель директора по научной работе Института прикладной математики им. М.В. Келдыша РАН, доктор физико-математических наук, профессор Г.Г. Малинецкий. Под его руководством разработан ряд компьютерных моделей для анализа, прогноза и мониторинга инновационных процессов в экономике России. Сегодня Россия находится в критическом положении, – ее главный ресурс состоит в изменении умов, а все техническое на сегодняшний день – второстепенно.