Учреждение образования
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

С. А. Барташевич

Моделирование систем обработки информации

Электронный курс лекций для студентов специальностей 1-36 06 01 «Полиграфическое оборудование и системы обработки информации», 1-40 01 02-03 «Информационные системы и технологии (полиграфический комплекс)»

Минск 2014

УДК 004.652(075.8)
ББК 32.973.я75
Б26

Рассмотрен и рекомендован редакционно-издательским советом Белорусского государственного технологического университета

Рецензенты:
кандидат технических наук, доцент кафедры информационных технологий автоматизированных систем БГУИР *А. М. Севернев*;
кандидат технических наук, доцент кафедры «Робототехнические системы» БНТУ *Р. В. Новичихин*

Барташевич, С. А.

Б26 Моделирование систем обработки информации : электронный курс лекций для студентов специальностей 1-36 06 01 «Полиграфическое оборудование и систем обработки информации», 1-40 01 02-03 «Информационные системы и технологии (издательско-полиграфический комплекс)» / С. А. Барташевич. – Минск : БГТУ, 2014. – 118 с.

В курсе лекций системно изложены теоретические основы обработки потоков информации и их управления методами и средствами построения и анализа моделей систем обработки информации для различных практических задач полиграфического комплекса.

Может быть использован студентами для подготовки к лекционным и практическим занятиям, экзамену или зачету и для выполнения контрольных работ.

УДК 004.652(075.8)
ББК 32.973.я75

Лекция 1. ПОНЯТИЕ МОДЕЛИ, ПРОЦЕССА МОДЕЛИРОВАНИЯ И ЦЕЛИ ЭТОГО ПРОЦЕССА

Модель – это совокупность физических или математических элементов и связей между ними, которые адекватно отображают определенные свойства объекта исследования или проектирования. Необходимо помнить, что, несмотря на адекватное отражение некоторых явлений действительности, модель не есть сама действительность. Обычно модель беднее объекта и лишь приближенно показывает некоторые исследуемые стороны и свойства объекта.

Моделирование предполагает построение действующей материальной, математической или виртуальной модели, обладающей свойствами, подобными (адекватными) свойствам рассматриваемой системы. Поэтому с помощью моделей можно имитировать, изучать функциональные системы и принимать решения относительно их наилучшего варианта, фактически не имея действующего образца.

Задача курса – изучить приемы и способы, необходимые для постановки задачи, формализации, изучения и интерпретации систем и процессов с целью их моделирования.

На практике исходным пунктом для моделирования является некоторая эмпирическая ситуация, выдвигающая перед инженером задачу, которую нужно решить. Прежде всего необходимо установить, в чем именно заключается «задача». Это связано с тем, что реальные ситуации редко бывают четко обозначенными и понятными, а сложное взаимодействие с окружающей средой часто делает точное описание таких случаев затруднительным. Процесс определения задачи, поддающейся математическому анализу, часто бывает продолжительным и требует владения новыми навыками, не имеющими прямого отношения к самому исследуемому явлению (например, беседы с математиками, чтение всевозможной литературы, имеющей отношение к данной области), – это все является элементами моделирования.

Часто, но не всегда, параллельно со стадией постановки задачи идет процесс выявления основных или существенных особенностей явления. Этот этап схематизации, упрощения (идеализации) играет решающую роль в переводе существенных факторов на язык математических понятий и величин. Как правило, это самая трудная стадия моделирования. После построения модели и формулировки ее качественной стороны исследователь приступает к количественному опи-

санию, которое заканчивается проверкой адекватности (соответствия) полученной модели опытным данным. В действительности адекватность модели до некоторой степени обычно проверяется в ходе постановки задачи. Уравнения или другие математические соотношения, сформулированные в модели, постоянно сопоставляются с исходной ситуацией.

Существует два основных требования проверки на адекватность. Во-первых, сама математическая основа модели (которая и составляет ее существо) должна не противоречить законам физики и подчиняться законам математической логики. Во-вторых, справедливость модели зависит от ее способности адекватно описывать экспериментальную ситуацию. Статистическая проверка адекватности осуществляется с использованием статистических критериев, например критерия Фишера. Для их определения применяются данные по воспроизводимости эксперимента и по отклонению экспериментальных величин от значений, предсказанных моделью. Методика расчета этих критериев описана в различных статистических руководствах.

Исходя из понятия полноты и адекватности отражения исследуемых процессов, модели можно разделить на упрощенные и сложные.

Приближенный результат, который получается на упрощенной модели быстрее, может оказаться более эффективным, чем результат усложненной модели, на получение которого уходит больше времени и средств. В этом заключается преимущество упрощенной модели. Недостаток же ее в том, что она, как правило, дает приближенную оценку процессов, происходящих в объекте, и не всегда позволяет выявлять их взаимовлияние.

Реальный объект бесконечен для познания, и поэтому никакая модель не может быть абсолютно идентичной исследуемому объекту или процессу. Поэтому моделирование проводят с помощью нескольких моделей, построение которых имеет своей целью разную детализацию процесса на различных этапах его изучения.

В общем случае различают два вида моделирования: физическое и математическое.

Физическое моделирование предполагает экспериментальное изучение различных физических явлений, основанное на их физическом подобии. Метод заключается в создании материальной экспериментальной модели в уменьшенном масштабе и проведении экспериментов. Полученные результаты анализируются и затем переносятся на явления в реальном масштабе. Физическое моделирование осуществляется, во-первых, когда натурные испытания очень

трудно выполнить из-за величины исследуемого объекта или его характеристик (давление, температура и т. д.). Во-вторых, когда наука не располагает точным математическим описанием моделируемого явления или такое описание слишком громоздко и требует больших вычислений.

Математическое моделирование – это абстрактный процесс, записанный в виде определенных символов с помощью системы математических соотношений. Поэтому такую модель называют формализованной (символьной). Абстрактной моделью в полиграфии является схема технологического процесса полиграфического производства или функциональные схемы систем обработки информации. Математическое моделирование гораздо дешевле физического, однако при проектировании, особенно технических объектов, математическое моделирование проверяется на макетах и экспериментальных образцах, т. е. оно не должно противоречить физическому объекту. По характеру отображаемых свойств математические модели делят на функциональные и структурные. Первые описывают функционирование моделируемого объекта и представляются уравнениями или системами уравнений. Вторые описывают только структурный состав объекта и представляются в виде матриц, операционных карт или таблиц.

Сам процесс моделирования состоит из трех стадий:

1. Формализация (переход от реального объекта к модели).

2. Моделирование (получение результатов путем решения уравнений или проведения экспериментов).

3. Интерпретация (перевод результатов моделирования в область реальности).

Когда исследование различных параметров в виде математической модели в силу сложности последней или необходимости выполнения больших вычислений производят на компьютере, говорят о компьютерном моделировании. Оно включает конструирование модели и постановку вычислительных экспериментов на этой модели с целью ее анализа и оптимизации или проектирования процессов и объектов. В этом случае математическая модель в соответствии с методом решения поставленных задач представляется в алгоритмической языковой форме. Специфика алгоритма состоит в отражении последовательности действий исследуемого процесса или явления в искусственной вычислительной среде, которой является компьютер. Достоинство компьютерного моделирования – это сочетание в себе как теории, так и эксперимента. Компьютерная модель дает

возможность относительно быстро и беззатратно исследовать ее свойства и поведение в любых возможных ситуациях. Причем вычислительные эксперименты на этих моделях позволяют их глубоко изучить и проанализировать.

При компьютерном моделировании для облегчения вычислительных задач математическая модель преобразовывается в алгоритмическую форму, и сформулированные по такому принципу задачи могут решаться как аналитическими методами, так и приближенными численными. В первом случае процесс компьютерного моделирования называется еще аналитическим.

В тех случаях, когда с помощью ЭВМ само явление, описанное математически, моделируется с сохранением логической структуры и последовательности его во времени, моделирование называется имитационным.

Аналитические модели в силу допускаемых упрощений для использования хорошо разработанного математического аппарата иногда ставят под сомнение точность результатов этого моделирования.

Имитационное моделирование осуществляется с целью воспроизведения исследуемой системы на основе результатов взаимосвязей между ее элементами. Обычно имитационные модели строятся для поиска оптимального решения в условиях ограниченных ресурсов, когда другие математические модели слишком сложны. В полиграфии примером имитационного моделирования является автоматизированная система обработки заказа. В этом случае технологический процесс представляется как упорядоченный набор работ, которые необходимо выполнить для изготовления заказываемых изданий. Такая система позволяет полиграфистам имитировать процесс изготовления полиграфической продукции с целью его оптимизации по времени и технологическим возможностям и определить стоимость заказа.

Моделирование – это сложный творческий процесс и его искусством могут овладеть только те, кто обладает оригинальным мышлением, изобретательностью и глубокими знаниями физических явлений, которые необходимо моделировать.

Проектирование – процесс создания модели объекта. Моделирование – это оценка результата проектирования.

Моделирование производится с целью:

– предсказания последствий изменения параметров системы или условий ее работы, когда такое изменение в реальной обстановке связано с большими материальными затратами или риском;

– изучения, переделки и усовершенствования существующих систем, познания и изучения систем, которые пока еще не существуют в реальной действительности;

– обучения персонала и демонстрации методов работы моделей и систем.

Таким образом, при изучении одного и того же явления могут применяться принципиально разные модели и в то же время одна и та же модель может описывать разные процессы. Поэтому для оценки моделей используются следующие критерии их эффективности:

1. Точность математической модели – это свойство, отражающее степень совпадения истинных значений параметров со значениями, рассчитанными с помощью модели. Однако при количественной оценке точности различных моделей необходимо учитывать следующие особенности этого сравнения:

− как правило, объекты и их модели характеризуются многими критериями, и для оценки точности таких моделей необходимо привести параметры моделей к одному или нескольким безразмерным показателям;

− в некоторых случаях модели составляются и используются для различных типов объектов и даже вариантов их использования. Например, математическая модель двухповодковой группы может применяться при анализе различных механизмов, содержащих эти двухповодковые группы, так как характер проявления свойств объекта и показатели точности этих свойств будут во многом определять условия функционирования модели. В результате оценка точности таких моделей варьируется в зависимости от характеристик моделируемых объектов;

− иногда результаты математического моделирования и оценка их точности сравниваются с экспериментально полученными значениями. Однако погрешности эксперимента во многих случаях оказываются соизмеримыми с погрешностями математического моделирования. Чтобы уменьшить влияние указанной погрешности, следует сравнивать результаты моделирования и эксперимента в некоторых так называемых стандартных ситуациях, тем самым преодолевая разночтения между условиями моделирования и эксперимента.

2. Экономичность математических моделей оценивается затратами машинного времени ЭВМ. Показателем экономичности модели можно считать число параметров, используемое в ней. Чем больше параметров, тем больше арифметических действий необходимо про-

изводить и больше оперативной памяти потребуется для расчета уравнения модели.

3. Степень универсальности математических моделей определяется их применимостью к анализу более или менее многочисленной группы в одном или многих режимах функционирования.

Анализируя приведенные основные требования, можно заметить их некоторую противоречивость, а именно: чем детальнее в модели отражаются различные закономерности процессов, тем она точнее и универсальнее, но и тем больше параметров в ней используется, а следовательно, и больший объем вычислений необходим для ее реализации.

Лекция 2. РЕГРЕССИЯ И ОСОБЕННОСТИ ЕЕ ИСПОЛЬЗОВАНИЯ ПРИ МОДЕЛИРОВАНИИ

Регрессия как метод моделирования позволяет получать функцию, приближенно описывающую случайное распределение некоторых статистических величин. Функциональную зависимость случайной выходной величины (y) от изменения входящих величин (x) называют регрессией, а модели, построенные на этой зависимости, – регрессионными.

В этом разделе мы изучим некоторые особенности работы с уравнениями регрессии, на которые студенты меньше всего обращают внимание.

К регрессионному анализу относятся проведение кривых через экспериментальные точки и подгонка (описание) экспериментальных данных какой-либо аппроксимирующей функцией, которая наилучшим способом описывает эти частные случаи. Это обычно осуществляется методом наименьших квадратов (МНК):

$$\sum_{i=1}^{n}\left(y_{Эi} - y_{Тi}\right)^2 = \min,$$

где n – число экспериментов; $y_{Эi}$ – выходные экспериментальные значения; $y_{Тi}$ – величины, полученные с помощью аппроксимирующей функции при подстановке в нее значений i-го опыта.

Широкому применению МНК способствует достаточно легкий (по сравнению с другими способами) аппарат расчета параметров регрессии. Наряду с МНК существуют и другие методы, позволяющие получить оценки параметров уравнений регрессии. Среди них наиболее известны метод наименьших модулей и метод минимакса (Чебышевское оценивание).

По первому методу необходимо, чтобы параметры регрессионного уравнения были найдены из соотношения минимизации суммы абсолютных отклонений:

$$\sum_{i=1}^{n}\left(y_{Э_i} - y_{Т_i}\right)^2 \to \min .$$

По второму – из требования минимизации максимального модуля отклонений между измеренными и вычисленными (теоретическими) значениями:

$$\max\left|y_{Э_i} - y_{Т_i}\right| = \min .$$

Упомянутым методам оценки параметров уравнения регрессии посвящена специальная литература, с которой вы можете ознакомиться. Мы же рассмотрим линейные по параметрам уравнения регрессии, которые наиболее часто используются при моделировании.

Эти модели являются самыми простыми, так как рассматривают объект как «черный ящик». Иными словами, никаких сведений о механизме, внутренних силах, причине и природе поведения изучаемого объекта они не дают. Однако это фактически самый простой и эффективный способ интерполирования. Эти линейные по параметрам модели используются для проведения оптимизации и выяснения влияния независимой переменной на зависимую. Однако при этом необходимо избегать неверной интерпретации самой регрессионной модели.

Пусть получено регрессионное уравнение следующего вида:

$$y = 10 + 50x_1 + 10x_2,$$

где x_1 – время, ч; x_2 – температура, °С.

Совершенно неправильно, исходя из величин коэффициентов приведенного уравнения, утверждать, что время оказывает большее влияние на процесс, чем фактор температуры.

Во-первых, температура и время измеряются в разных физических единицах и их сравнивать нельзя.

Во-вторых, если измерение времени привести не в часах, а в минутах, то уравнение примет вид

$$y = 10 + (50/60)x_1 + 10x_2.$$

В этом случае коэффициент перед x_1 меньше, чем перед x_2. Таким образом, только знак перед независимой переменной является той информацией, которой можно доверять.

По степени информированности исследователя об объекте существует деление объектов на три типа «ящиков».

Однако вернемся к примерам построения регрессионной модели.

Для построения регрессионной функции ($y_\text{Т}$) используется МНК, в соответствии с которым сумма квадратов отклонений (разность между экспериментальными и теоретическими значениями) должна быть минимальна:

$$\Phi = \sum_{i=1}^{n} \left(y_{\text{Э}_i} - y_{\text{Т}_i} \right) \to \min. \qquad (2.1)$$

Этот метод позволяет найти коэффициенты регрессионного уравнения, которое в общем виде описывается зависимостью y_T от значения k независимых факторов и имеет вид

$$y_{\text{Т}} = b_0 x_1 + \ldots + b_k x_k + b_{k+1}\left(x_1\right)^2 + b_{k+2}\left(x_2\right)^2 + \ldots \qquad (2.2)$$

Коэффициенты b_0, b_1, b_2, … находятся из выражения (2.1) путем минимизации функции (Ф), условием которого является равенство нулю всех частных производных:

$$2\sum_{i=1}^{n}\left(y_{\text{Э}_i} - y_{\text{Т}_i}\right)\frac{\partial y_{\text{Э}_i}}{\partial b_0} = 0,\ 2\sum_{i=1}^{n}\left(y_{\text{Э}_i} - y_{\text{Т}_i}\right)\frac{\partial y_{\text{Э}_i}}{\partial b_1} = 0,\ \ldots,$$

откуда получим систему уравнений, называемую нормальной системой МНК, для определения коэффициентов b_0, b_1, b_2 искомой функции $y_{\text{Т}}$:

$$\begin{cases} \sum_{i=1}^{n} y_{\text{Э}_i}\dfrac{\partial y_{\text{Э}_i}}{\partial b_0} - \sum_{i=1}^{n} y_{\text{Т}_i}\dfrac{\partial y_{\text{Т}_i}}{\partial b_0} = 0, \\ \sum_{i=1}^{n} y_{\text{Э}_i}\dfrac{\partial y_{\text{Э}_i}}{\partial b_1} - \sum_{i=1}^{n} y_{\text{Т}_i}\dfrac{\partial y_{\text{Т}_i}}{\partial b_1} = 0. \end{cases}$$

Решение данной системы рассмотрим на следующем примере. Необходимо аппроксимировать экспериментальную зависимость коэффициента насыщения бумаги K от толщины слоя краски на печатной форме h, значения которых приведены в таблице.

i	1	2	3	4	5	6	7
h_i	0,9	1,2	1,3	1,6	1,7	2,0	2,2
K_i	0,35	0,45	0,50	0,60	0,73	0,86	1,00

В приведенном примере измеряемая величина K_i зависит от одного фактора h_i, т. е. можно построить аппроксимирующую функцию $K(h)$.

Графическое представление табличных точек позволяет сделать вывод, что экспериментальные значения описываются простейшей зависимостью – линейной функцией. На основании сделанного предположения система уравнений, где в качестве входных и выходных величин выступают h_i и K_i соответственно, нормальная система МНК будет иметь вид

$$\begin{cases} 7b_0 + 10,9b_1 = 4,49, \\ 10,9b_0 + 18,23b_1 = 7,626. \end{cases}$$

В результате решения вышеприведенного уравнения определим его коэффициенты $b_0 \approx -0,144$ и $b_1 \approx 0,505$. Таким образом, линейная регрессия будет представлена функцией $K(h) = -0,144 + 0,505h$.

Для более точной аппроксимации кроме линейной могут использоваться параболические и другие стандартные функции. В случае параболической функции уравнение (2.2) примет вид

$$y_{\text{Т}} = b_0 + b_1 x + b_2 x^2,$$

а система уравнений МНК будет включать три уравнения:

$$\begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_1^2 = \sum\limits_{i=1}^{n} y_i, \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 + b_2 \sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} y_i x_i, \\ b_0 \sum\limits_{i=1}^{n} x_i^2 + b_1 \sum\limits_{i=1}^{n} x_i^3 + b_2 \sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} y_i x_i^2. \end{cases}$$

Если исследуемый процесс зависит от двух факторов x_1 и x_2, то уравнение (2.2) будет иметь вид

$$y_{\text{Т}} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2.$$

В этом случае нормальная система состоит из четырех уравнений, решение которых рассмотрено в специальной литературе, например [1].

К линейной множественной модели приводятся многие нелинейные модели подстановками и переобозначениями.

Сначала рассмотрим, как это можно сделать, а затем поясним нежелательные последствия, которые могут возникнуть из-за таких преобразований. В качестве примера обычно приводят уравнение следующего вида:

$$y = k \exp(-ax),$$

которое после логарифмирования можно записать так:

$$\ln y = \ln k - ax.$$

Заменяя переменные, получим $Z = b - ax$, где $\ln k = b$, $\ln y = Z$.

Рассмотрим еще один пример.

Уравнение вида $y = ax_1 / (1 + bx_2 + cx_3)^2$ с помощью

$$Z = \sqrt{\frac{x_1}{y}};\ d = \left(\frac{1}{\sqrt{a}}\right);\ f = \left(\frac{b}{\sqrt{a}}\right);\ g = \left(\frac{c}{\sqrt{a}}\right)$$

можно преобразовать в линейное уравнение:

$$Z = d + fx_2 + gx_3.$$

Существуют модели, преобразовать которые в линейные по приведенному подходу невозможно, например $y = a - b^c$.

В этом случае линеаризацию можно провести путем разложения данного уравнения в ряд Тейлора.

К требованиям, которые предъявляют к оценкам параметров регрессионного уравнения, относится получение несмещенных оценок, т. е. таких, при которых их выборочная (вычисленная) оценка будет максимально близкой к истинной. Однако линеаризация не всегда позволяет получать такие оценки. Рассмотрим ранее линеаризованные уравнения

$$\ln y = k - ax .$$

Используя метод наименьших квадратов, мы будем минимизировать логарифмы величин вместо самих величин. Другими словами, параметры для исходного уравнения заменяются параметрами линеаризированного уравнения, что, в общем-то, не совсем верно, хотя ее можно использовать как предварительную.

В общем случае нахождение регрессионной зависимости для экспериментальных данных включает два этапа:

1) выбор вида аппроксимирующей функции y_T;

2) определение значения ее коэффициентов.

Первый этап удобно проводить, основываясь на графическом представлении табличных данных или исходя из известных теоретических представлений об исследуемом процессе.

Если график однофакторной регрессии близок к прямой линии или параболе, то для определения параметров выбранной функции применяют метод наименьших квадратов.

Если для аппроксимации выбрана известная стандартная математическая функция (экспоненциальная, логарифмическая, показательная и др.), то предварительно удобно провести соответствующую замену координат, чтобы привести зависимость к линейной. В этом случае нормальная система метода наименьших квадратов будет иметь самый простой вид.

Для ответа на вопрос, какая из функций подходит наилучшим образом, кроме графического метода, можно воспользоваться условиями постоянства соответствующих разделенных разностей.

Регрессионный и корреляционный анализ

Построение математической модели на основе экспериментальных данных, которые, как уже говорилось, являются случайными, не ограничивается отысканием коэффициентов регрессионного уравне-

ния. Обработка опытных данных на основе статистических методов включает следующие этапы:

1) оценка качества проведения опытов;

2) нахождение и проверка значимости коэффициентов уравнения регрессии;

3) оценка адекватности полученного уравнения исследуемому процессу.

Такое исследование называется регрессионным анализом.

При этом принимаются следующие допущения:

– входные параметры x_i измеряются с пренебрежимо малой ошибкой. Появление ошибки в определении y_i объясняется наличием в процессе воздействий, которые не учтены в уравнении регрессии;

– результаты наблюдений над выходной величиной y_1, y_2,..., y_n представляют собой независимые нормально распределенные случайные величины;

– при проведении n опытов каждый из них должен быть перепроверен m раз, то есть говорят, что проведено m параллельных опытов с одними и теми же значениями исходных факторов.

Качество проведения эксперимента определяется в ходе проверки однородности дисперсии. Пусть y_{ij} – результат, полученный в i-м опыте в ходе его j-й проверки (j-й параллельный опыт), тогда среднее по параллельным опытам значение i-го эксперимента равно

$$\overline{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$$

и выборочная дисперсия, характеризующая разброс опытных значений, найдется по формуле

$$D_i = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2 .$$

Далее, для всех опытов $i = 1, 2, \ldots, n$ определяются максимальное значение дисперсии и сумма этих дисперсий для расчета критерия Кохрена:

$$G = \frac{D_{\max}}{\sum_{i=1}^{n} D_i}. \qquad (2.3)$$

Это значение сравнивается с табличным критерием Кохрена: $G_p(n, m-1)$, где n и m называются степенями свободы; p – выбран-

ный уровень значимости (характеризует вероятность ошибки), обычно $p = 0{,}05$. Если найденное значение G меньше, чем табличное ($G < G_p$), то дисперсии считаются однородными.

Далее рассчитывается среднее значение дисперсий – дисперсия воспроизводимости:

$$D_{\text{восп}} = \frac{1}{n}\sum_{i=1}^{n} D_i , \tag{2.4}$$

которая необходима на этапе проверки значимости найденных по методу наименьших квадратов коэффициентов регрессии.

Для проверки значимости вычисляются расчетные значения критерия Стьюдента для каждого коэффициента b_j , $j = 0, 1, \ldots, l,$ где l определяет количество корреляционных коэффициентов в регрессионном уравнении. Для случая линейного уравнения с одним изменяющимся параметром $l = 2$.

$$t_j = \left| b_j \right| / S_{bj},$$

где S_{bj} – среднеквадратичная ошибка в определении j-го коэффициента. Например, для линейной регрессии от одного фактора ее можно найти по следующим формулам:

$$S_{b_0}^2 = \frac{D_{\text{восп}} \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}; \quad S_{b_1}^2 = \frac{nD_{\text{восп}}}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Найденные значения критерия сравниваются с табличными значениями $t_p(f)$, где f – число степеней свободы, $f = n(m - 1)$.

Если условие $t_j > t_p(f)$ выполняется, то коэффициент b_j является значимым (существенно отличается от нуля), если нет – коэффициент b_j незначим и его следует исключить из уравнения. Для нового вида регрессионной функции коэффициенты пересчитываются заново.

Адекватность уравнения проверяется по критерию Фишера:

$$F = D_{\text{ад}} / D_{\text{восп}}, \tag{2.5}$$

для которого дисперсия воспроизводимости вычисляется по (2.2), а дисперсия адекватности – по следующей формуле:

$$D_{\text{ад}} = \frac{m}{n-k}\sum_{i=1}^{n}\left(f_i - \overline{y_i}\right)^2 ,$$

где k – окончательное число коэффициентов регрессии после исключения незначимых; f_i – значение, рассчитанное подстановкой в уравнение регрессии факторов i-го опыта (теоретическое значение); $\overline{y_i}$ – среднее значение i-го опыта (экспериментальное значение).

Расчетное значение критерия сравнивается с табличным для уровня значимости p и чисел степеней свободы $n(m-1)$ и $n-k$.

Если выполняется условие $F < F_p$ ($n(m-1)$, $n-k$), то уравнение адекватно процессу и может использоваться для аппроксимации регрессии. Если нет, то уравнение с недостаточной степенью точности описывает процесс, тогда необходимо перейти к более сложной математической модели или провести эксперименты с меньшим интервалом изменения факторов.

Проверка на адекватность возможна, если существуют степени свободы критерия Фишера, то есть когда $k < n$. Если же число опытов меньше числа коэффициентов, то следует построить и проверить на адекватность более простую модель.

При отсутствии параллельных опытов ($m = 1$) качество эксперимента можно оценить не по дисперсии воспроизводимости, а по дисперсии относительного среднего:

$$D_y = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}{n-1},$$

а вместо дисперсии адекватности использовать остаточную дисперсию:

$$D_y = \frac{\sum_{i=1}^{n}\left(f_i - \overline{y}\right)^2}{n-k}.$$

Тогда (2.3) примет вид

$$F = D_{\text{ост}} / D_y .$$

Часто для нахождения дисперсии воспроизводимости проводят параллельные опыты не для всех, а только для одного набора факторов, считая, что точность проведения всех опытов одинаковая. Пусть q – номер этого одного проверяемого опыта, тогда в качестве дисперсии воспроизводимости берут одну выборочную дисперсию, то есть

$$D_{\text{восп}} = \frac{\sum_{i=1}^{m}\left(y_{qi} - \overline{y_q}\right)^2}{m-1}.$$

Критерий Фишера в этом случае примет вид

$$F = D_{\text{ост}} - D_{\text{восп}}.$$

В большинстве случаев реальные экспериментально исследуемые процессы зависят от многих факторов. Большое число исходных переменных затрудняет проведение опытов, построение математических моделей и решение с их помощью задачи оптимизации. Если же параметры связаны между собой, то можно сократить их число, вычисляя парные коэффициенты корреляции.

В общем случае если две случайные величины связаны друг с другом, то имеет место корреляционная зависимость, и в этом случае коэффициент корреляции применяется для оценки этой связи.

Для вычисления коэффициента корреляции используются следующие выборочные характеристики:

1) выборочное среднее

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k ;$$

2) выборочная дисперсия

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2 .$$

где $S = \sqrt{S^2}$ обозначает выборочное среднеквадратичное отклонение.

Формула для расчета выборочного коэффициента корреляции имеет вид

$$r_{x,y} = \frac{\sum_{i=1}^{n}\left(\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right)}{\left(n-1\right)S_x S_y}.$$

Коэффициент может принимать значения от –1 до 1. Близость $r_{x,y}$ свидетельствует о линейной связи между x и y. Если коэффициент очень мал, то величины не закоррелированы (не взаимодействуют). Однако надо помнить, что это справедливо для нормально распределенных случайных величин. В противном случае нулевое значение коэффициента корреляции говорит не о независимости величин, а только об отсутствии линейной связи между ними.

Лекция 3. ПРИНЦИПЫ ПОСТРОЕНИЯ МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ РАСПОЗНАВАНИЯ ОБРАЗОВ

Длительное время методы распознавания использовались в таких плохо формализованных областях, как медицина, биология, социология, психология и т. д. При этом применялись в основном качественные показатели, не позволяющие точно описать соответствующие явления. Если и получались числовые характеристики, то они, как правило, были основаны на работе органов таких чувств, как зрение, слух, осязание.

На начальных этапах для построения алгоритмов распознавания использовалось классическая теория статистических решений, основанная на экспериментальных исследованиях и не имеющая математического обоснования, что позволяло обеспечить лишь определение класса, к которому может быть отнесен исследуемый объект. Приведем примеры, поясняющие вышесказанное.

Так, в медицине на основании ряда косвенных анализов возникает возможность распознать заболевание, т. е. отнести его к одному из известных. В метеорологии по накопленным результатам о давлении воздуха, температуре различных слоев атмосферы, скорости движения воздушных потоков можно дать прогноз погоды. В химии, пользуясь спектральным анализом, можно отнести вещество к тому или иному классу химических соединений. Данный метод используется в геологии при прогнозировании возможности залегания новых месторождений по данным геохимического и геофизического анализа, а также в таких областях, как экономика, психология, лингвистика.

В настоящее время математический аппарат, привлекаемый для решения задач распознавания, существенно расширился за счет использования методов алгебры логики, некоторых разделов математического программирования, а также нейросетевых технологий. Это не только породило множество алгоритмов распознавания, но и привело к возникновению значительного количества новых методов распознавания.

Распознавание представляет собой задачу преобразования входной информации. Рассмотрим задачи, возникающие при построении системы распознавания.

Первая из них заключается в определении полного перечня признаков, так называемого словаря признаков, для априорного описания

классов. Признаки могут быть подразделены на детерминированные, вероятностные, логические, структурные.

Детерминированные – имеют конкретные числовые значения и могут рассматриваться в качестве координат точки в признаковом пространстве, соответствующем данному объекту. Например, размер знака, координаты его расположения, оптическая плотность.

Вероятностные – принимают случайные значения, например, уровни шума при диагностике работающей машины или признаки знаков при рукописном их написании, когда начертание знаков зависит от личности человека.

Логические – это элементарные высказывания, принимающие два значения («да», «нет» или «истина», «ложь») с полной определенностью и не имеющие количественного выражения. В качестве логических признаков можно рассматривать наличие некоторых свойств у распознаваемых объектов, например, засечек у некоторых гарнитур или наличие увлажняющего или фальцевального аппаратов в исследуемой полиграфической машине, а также признаки, у которых важна не величина признака, а лишь факт его наличия или отсутствия. На практике логические признаки используются тогда, когда ошибками измерения можно пренебречь или интервалы значений признаков выбраны так, что ошибки измерений практически не влияют на достоверность принимаемых решений. Так, например, при техническом диагностировании машины (станка) решение о выходе его из строя принимается лишь тогда, когда фактические значения определенных параметров (признаков) выходят за пределы признанных интервалов.

Структурные – представляют собой элементы структуры объекта, иначе называемые терминалами. Каждый объект может рассматриваться как цепочка терминала. Например, слово имеет набор знаков, знак – набор линий или примитивов и т. д.

Вторая задача – описание классов распознаваемых объектов или явлений, составление априорного алфавита класса. Основное в данной задаче – выбор надлежащего принципа классификаций. При решении этой задачи необходимо каждому классу назначить соответствующие числовые параметры детерминированных и вероятностных признаков и определить значения логических признаков. Причем в зависимости от объема исходной априорной информации могут быть использованы методы непосредственной обработки и назначения числовых значений признаков или их определения,

а также формирования алфавита классов в результате обучения или самообучения.

Следующий этап связан с разработкой априорного словаря признаков. Он формируется на основе результатов решения первой задачи. Необходимо также описание всех классов априорного алфавита классов на языке признаков, включенных в априорный словарь признаков. Такая задача не имеет однозначного решения. Так, если признаки детерминированные, то описаниями классовых объектов на языке этих признаков являются их эталоны.

Если признаки распознаваемых объектов логические, имеющие количественное выражение, то для описания классов состояния объекта на языке признаков необходимо определить диапазоны значений признаков, соответствующие этим классам.

Если признаки распознаваемых объектов структурные, то описанием классов объектов являются языки, состоящие из предложений, каждое из которых характеризует особенности объектов, принадлежащих исключительно одному из классов. Априорные пространства признаков разбиваются на области, соответствующие классам априорного алфавита. Это разбиение должно быть выполнено таким образом, чтобы обеспечивались минимальные значения ошибок отнесения классифицируемых состояний объекта к чужим классам.

Третья, ключевая, ***задача*** – выбор алгоритмов распознавания, обеспечивающих отнесение распознаваемого объекта или явления к тому или другому классу или к их некоторой совокупности. Алгоритмы распознавания основываются на сравнении той или другой меры сходства состояния распознаваемого объекта с каждым из имеющихся классов.

В большинстве алгоритмов многомерной классификации используется понятие, которое носит название «мера сходства» или «мера подобия» между объектами. В практической работе применяются три меры сходства: показатели расстояния, коэффициент корреляции, коэффициент подобия.

Наиболее простая и понятная мера сходства – расстояние между объектами. Для двух- и трехмерного пространства признаков она имеет простой и наглядный вид. Предположим, у нас имеется два объекта А и В (рис. 3.1), каждый из которых характеризуется набором из двух признаков X_1^1, X_2^1 для первого объекта и X_1^2, X_2^2 для второго объекта. Здесь верхние индексы – это номер объекта, а нижние – это свойства этих объектов в координатах X_1 и X_2.

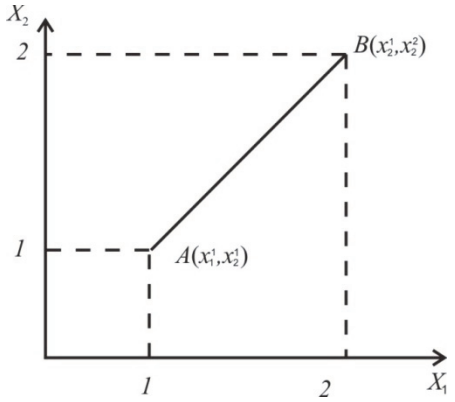


Рис. 3.1. Расстояние между объектами как мера сходства

Расстояние между двумя точками по прямой называется евклидовым расстоянием. Способ расчета такого расстояния известен нам из школьной математики. При переходе к пространству более высокой размерности расстояние между объектами *i* и *j* рассчитывается по формуле

$$\alpha_{ij} = \sqrt{\sum_{k=1}^{m} (X_i^k - X_j^k)^2},$$

где *m* – число признаков.

Эта мера близости еще называется среднеквадратичным расстоянием между сравниваемыми объектами.

Другой наиболее часто употребляемой мерой расстояния является расстояние по Хэммингу:

$$\alpha_{ij} = \sum_{k=1}^{m} \left| X_i - X_j \right|.$$

Существуют и более сложные метрики расстояния, определяющие меру сходства. Рассчитанные по Евклиду и Хэммингу расстояния для вышеприведенного примера будут равны соответственно:

$$d_E = \sqrt{(3-1)^2 + (3-1)^2} = 2{,}83; \; d_x = |3-1| + |3-1| = 4.$$

Использование коэффициента корреляции поясним следующим примером. Имеется географическая карта, на которую нанесены изолинии двух явлений высоты и заселенности. Если изолинии совпадают, то связь между ними полная; если они пересекаются под прямым углом, то связь отсутствует. Таким образом, корреляцию можно выразить с помощью геометрической интерпретации. На рис. 3.2 изображены вектора $\overline{h}_1$ и $\overline{h}_2$, угол между которыми равен α. Связь между этими векторами определяется через косинус угла между ними.

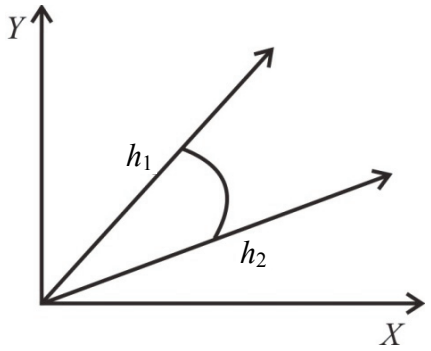


Рис. 3.2. Геометрическая интерпретация корреляционной зависимости

Здесь же виден и определенный недостаток этой меры близости: если длины векторов сильно различаются, а угол мал, то сходство достаточно сомнительное.

Коэффициенты корреляции применяются в методах распознавания, которые базируются на теории факторного анализа.

Коэффициенты подобия используются для описания объектов, признаки которых имеют дихотомические значения (0 или 1). Формулы для их вычисления выглядят следующим образом:

$$s_{ij} = \frac{n_{ij}^{(1,1)}}{n_i^{(1)} + n_j^{(1)} - n_{ij}^{(1,1)}},\ (0 \le s_{ij} \le 1);\ s_{ij} = \frac{p_{ij}}{m},\quad (0 \le s_{ij} \le 1);$$

$$s_{ij} = \frac{n_{ij}^{(1,1)}}{m},\ (0 \le s_{ij} \le 1);\ s_{ij} = \frac{p_{ij} - q_{ij}}{m},\quad (0 \le s_{ij} \le 1);$$

$$s_{ij} = \frac{n_{ij}^{(1,1)}}{n_{ij}^{(1,1)} + q_{ij}},\quad (0 \le s_{ij} \le 1);\ s_{ij} = \frac{2n_{ij}^{(1,1)}}{2n_{ij}^{(1,1)} + q_{ij}},\quad (0 \le s_{ij} \le 1);$$

$$s_{ij} = \frac{p_{ij}}{2m - p_{ij}} = \frac{p_{ij}}{m + q_{ij}},\ (0 \le s_{ij} \le 1),$$

где $n_{ij}^{(1,1)}$– число совпадений признаков со значением 1; $n_i^{(1)}$, $n_j^{(1)}$– число единичных признаков у i-го и j-го объектов соответственно; p_{ij} – общее число совпадающих признаков; q_{ij} – общее число несовпадающих признаков; m – общее число признаков.

Если признаки являются дихотомическими, то определение меры сходства можно проводить по косинусу угла, используя формулу

$$\cos(X,Y) = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2} \cdot \sqrt{\sum Y_i^2}}.$$

Рассмотрим пример расчета, исходные данные приведены в таблице.

Объект	Признак						
	1	2	3	4	5	6	7
1	0	0	0	1	0	1	1
2	0	1	0	0	1	1	0
3	1	1	1	1	0	0	1

Как следует из таблицы:

$\cos(1, 2) = (0 + 0 + 0 + 0 + 0 + 1 + 0) / (1 + 1 + 1)^{0,5} \cdot (1 + 1 + 1)^{0,5} = 0,333$;

$\cos(1, 3) = (0 + 0 + 0 + 1 + 0 + 0 + 1) / (1 + 1 + 1)^{0,5} \cdot (1 + 1 + 1 + 1 + 1)^{0,5} = 0,13$;

$\cos(2, 3) = (0 + 1 + 0 + 0 + 0 + 0 + 0) / (1 + 1 + 1)^{0,5} \cdot (1 + 1 + 1 + 1 + 1)^{0,5} = 0,067$.

Таким образом, наибольшее сходство имеется между объектами 1 и 2, наименьшее – между 2 и 3.

Выбор метода подобия определяется характером решаемой задачи и во многом произволен.

Рассмотренные выше меры сходства, в сущности, являются эвристическими и иногда могут противоречить друг другу. Некоторым преимуществом обладают меры сходства, основанные на понятии, связанном с количеством информации. Этот вопрос будет изложен более подробно в лекциях 15–16.

Четвертая задача – разработка алгоритма управления работой системы распознавания. Его назначение в том, чтобы функционирование системы распознавания было оптимальным и выбранный критерий качества этого процесса достигал экстремального значения. В качестве такого критерия может использоваться, например, вероятность правильного решения задачи, среднее время ее решения, расходы на ее решение, и, наконец, должен быть выбран показатель или система показателей эффективности системы распознавания и оценки их значений. Оценка значений выбранных показателей эффективности производится на основе экспериментальных исследований реальной системы распознавания или ее модели (физической или математической).

С учетом вышесказанного системы распознавания можно подразделить на простые и сложные, без обучения и с обучением. Также при создании моделей и алгоритмов распознавания могут использоваться различные методы: детерминированные, вероятностные, логические и т. д.

Простые системы, как правило, применяют для распознавания физически однородной информации. Например, автоматы, использующие в качестве признака вес жетона или монеты; автоматы для обработки деталей, в которых в качестве признаков для описания классов применяются линейные размеры. *Сложные системы* могут использовать физически неоднородную информацию. Например, система медицинской диагностики, которая применяют в качестве признаков температуру, динамику кровяного давления, состав крови, кардиограмму и т. п.; системы, предназначенные для распознавания военной техники вероятного противника; встроенные системы мониторинга сложного технологического оборудования и его управления и т. д.

Сложные системы по способу получения информации о признаках распознаваемых объектов или явлений можно подразделить на одноуровневые и многоуровневые. В *одноуровневых системах* информацию о признаках распознаваемых объектов получают в результате непосредственной обработки прямых измерений используемых технических средств (датчики перемещений, температуры, давления и т. д.).

В *многоуровневых системах* информацию о признаках получают на основе косвенных измерений, для которых используются специализированные локальные распознающие системы. Полученные признаки называются первичными или признаками первого уровня. Они используются распознающими устройствами второго уровня в качестве исходной информации для получения признаков второго уровня. Признаки второго уровня, в свою очередь, используются для получения признаков третьего уровня и т. д. К последней группе относятся признаки, непосредственно используемые в процессе распознавания, т. е. признаки, входящие в рабочий словарь признаков системы распознавания.

Если классифицировать системы распознавания по объему и способу использования первоначальной априорной информации об объектах, то и простые, и сложные системы можно разделить на системы без обучения, обучающиеся и самообучающиеся.

В *системах без обучения* априорной информации достаточно для того, чтобы в соответствии с выбранным принципом классификации разделить все множество состояний объектов на классы, составить словарь признаков и на основе непосредственной обработки исходных данных описать каждый класс объектов на языке этих признаков.

В *обучающихся системах* априорной информации достаточно лишь для того, чтобы в соответствии с заданным принципом классификации разделить все множество состояний объектов на классы и со-

ставить словарь признаков. Однако этой информации недостаточно для описания самих классов на языке признаков. Но на основе исходных сведений можно сформулировать обучающие последовательности, с помощью которых можно организовать процесс обучения. Цель обучения – найти разделяющие функции путем многократного предъявления системе распознавания различных наборов данных, описывающих состояния исследуемых объектов, с указанием классов, к которым эти данные принадлежат. После обучения систему распознавания необходимо проверить (проэкзаменовать), корректируя полученные результаты до тех пор, пока количество ошибок в среднем не достигнет определенного уровня.

В *самообучающихся системах* априорной информации достаточно лишь для определения словаря признаков, но недостаточно для проведения классификации. На стадии обучения системе предъявляют исходную совокупность состояний объектов, заданных значениями своих признаков, но из-за ограниченного объема первоначальной информации система не получает указаний о том, к какому классу принадлежат исходные наборы признаков. Эти указания заменяются набором правил, в соответствии с которыми на стадии самообучения система распознавания сама вырабатывает набор решающих правил и классификацию, которая может отличаться от общепринятой, и в дальнейшем ее придерживается.

Цель обучения или самообучения – получить такое количество информации, которого достаточно для функционирования системы распознавания.

В настоящее время получили широкое распространение обучающиеся и самообучающиеся системы распознавания на основе нейронных сетей, хорошо зарекомендовавшие себя в условиях отсутствия полной первоначальной априорной информации.

Если классифицировать системы по характеру признаков, то можно и их подразделить на логические, вероятностные, детерминированные, структурные и комбинированные.

В *логических системах* для построения алгоритмов распознавания используются логические методы, основанные на дискретном анализе и булевой алгебре. Применение подобных методов предусматривает формализованные логические связи в описании объекта, в которых переменные – логические признаки распознаваемых объектов или их состояний, а неизвестные величины – классы, к которым эти объекты (состояния) относятся.

В *вероятностных системах* используются вероятностные методы распознавания, основанные на теории статистических решений. Применение этих методов предусматривает вероятностные зависимости между признаками распознаваемых объектов и классами, к которым эти объекты относятся.

Детерминированные системы – это системы с применением детерминированных методов распознавания, которые предусматривают координаты эталонов классов в признаковом пространстве либо координаты объектов, принадлежащих к соответствующим классам.

Комбинированные системы – это системы, которые для построения алгоритмов применяют специальный разработанный метод вычисления оценок АВО – алгоритм вычисления оценок. При этом используются таблицы, где содержатся объекты, принадлежащие соответствующим классам, а также значения признаков, которыми характеризуются эти объекты.

Структурные системы – это системы, применяющие структурные методы распознавания, использование которых требует совокупностей предложений, описывающих все множество объектов, принадлежащих всем классам алфавита классов системы распознавания.

Построение и функционирование систем распознавания связано с накоплением и анализом больших объемов априорной информации. Для диагностики полиграфического оборудования могут быть использованы в зависимости от сложности поставленных задач как простые и сложные системы распознавания без обучения, так и системы с обучением.

Лекция 4–5. ИСПОЛЬЗОВАНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ МОДЕЛИРОВАНИЯ В ПОЛИГРАФИИ

4.1. Задачи искусственных нейронных сетей

Искусственные нейронные сети (ИНС) получают все большее распространение. В последние годы разработки в этой области представляют большой интерес не только для теоретиков, работающих в области искусственного интеллекта, но и для инженеров. Областей применения ИНС множество.

Нейросетевой подход показал свою эффективность при решении следующих задач в полиграфии:

1. Распознавание образов. Задача состоит в отнесении входного набора данных, представляющего распознаваемый объект, к одному из заранее известных классов. В число этих задач входит распознавание рукописных и печатных символов при оптическом их вводе в ЭВМ.

Оптическое распознавание печатного текста – частная задача более сложной проблемы распознавания образов, которая реализуется с помощью нейронных сетей, в частности сети Хопфилда. На рынке услуг, связанных с распознаванием печатного и рукописного текстов, представлены многие пакеты программ, среди которых самым распространенным и широко используемым является ABBYYFineReader. Алгоритм работы указанного пакета основан на сочетании шаблонного и структурного методов распознавания образов. При анализе образа выделяются реперные точки объекта, так называемые пятна. В качестве пятен могут выступать концы линий, узлы, места изломов линий, места пересечения линий, крайние точки. После выделения пятен определяются связи между ними – отрезок, дуга. Таким образом, итоговое описание представляет собой граф, который и сравнивается с эталоном.

В результате устанавливается соответствие между реперными точками образца и эталона, после чего находится степень деформации связей и определяется правильность распознаваемого символа. При большой деформации между реперными точками программа может выполнять дополнительную коррекцию, позволяющую увеличить качество распознавания спорных символов. Это производится на основе анализа буквосочетаний, грамматического анализа и других методов.

2. Кластеризация данных. Задача состоит в группировке входных данных по присущему им сходству. Алгоритм определения признаков сходства данных (определение расстояния между векторами, вычисление коэффициента корреляции и другие способы обучения) закладывается в нейросеть при ее построении и обучении. Сеть кластеризует данные на заранее неустановленное число кластеров. Наиболее известные применения кластеризации связаны со сжатием данных, анализом данных и поиском в них закономерностей.

3. Аппроксимация функций. Имеется набор экспериментальных данных, представляющий значение Y_i неизвестной функции от аргумента X_i, где $i = 1, ..., n$. Требуется найти аппроксимирующую функцию, удовлетворяющую некоторым критериям. Эта задача актуальна при моделировании сложных систем и создании систем управления сложными динамическими объектами.

4. Предсказания. Имеется набор $Y(t_1)$, $Y(t_2)$, …, $Y(t_n)$. Значение Y представляет поведение системы в моменты времени t_1, t_2, ..., t_n. Требуется по предыдущему поведению системы предсказать ее поведение в момент времени t_{n+1}. Эта задача актуальна для технической диагностики, для систем принятия решений.

Особенности построения нейронных сетей. Основу каждой нейронной сети составляют относительно простые, в большинстве случаев однотипные элементы, имитирующие работу нейронов мозга.

Таким образом, нейронная сеть – это сеть с конечным числом слоев из однотипных элементов – аналогов нейронов человеческого мозга с различными типами связей между слоями. При этом число элементов в слоях выбирается исходя из сложности решаемой задачи, а число слоев берется как можно меньше для сокращения времени вычислений.

Математическая модель нейрона представлена на рис. 4.1:

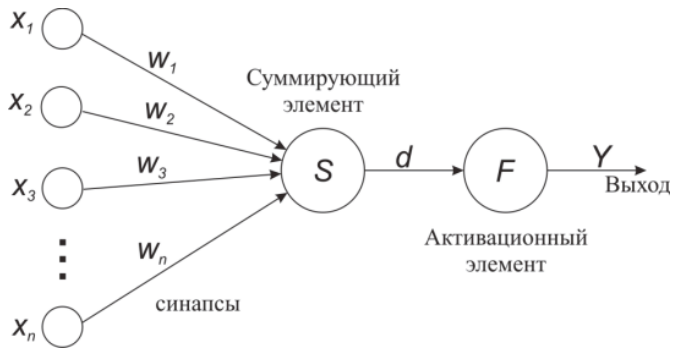


Рис. 4.1. Математическая модель нейрона

Биологический нейрон моделируется как устройство, имеющее несколько входов (дендриты) и один выход (аксон). Нейрон характеризуется своим текущим состоянием по аналогии с нервными клетками головного мозга, которые могут быть возбуждены или заторможены. Он обладает группой синапсов – одновременных входных связей, соединенных с выходами других нейронов, а также имеет аксон – выходную связь, с которой сигнал поступает на синапсы следующих нейронов.

Каждому входу ставится в соответствие некоторый весовой коэффициент w_i, характеризующий пропускную способность канала и определяющий степень влияния сигнала с этого входа на сигнал на выходе. Входные сигналы умножаются на соответствующие весовые коэффициенты, все произведения суммируются, определяя уровень активации нейрона. Нейрон осуществляет взвешенное суммирование входных воздействий, и далее это значение является аргументом активационной функции нейрона. В зависимости от конкретной реализации обрабатываемые нейроном сигналы могут быть аналоговыми или цифровыми.

Здесь множество входных сигналов обозначено вектором $\overline{X}$. Совокупность весовых коэффициентов обозначена вектором $\overline{W}$. Нейрон состоит из взвешенного сумматора и нелинейного элемента. Сумматор (аналог биологического нейрона) складывает взвешенные входы алгебраически:

$$d = \sum_{i=1}^{n} x_i w_i ,$$

где x_i – входные сигналы; w_i – весовые коэффициенты.

Выход нейрона является функцией его состояния $Y = F(d)$, где F – нелинейная функция, называемая функцией активации.

4.2. Основные функции активации

Функция активации определяет выходной сигнал нейрона. Если функция активации одна и та же для всех нейронов, сеть называют однородной (гомогенной). Если же активационная функция зависит еще от одного или нескольких параметров значения, которые меняются от нейрона к нейрону, то сеть называют неоднородной (гетерогенной). На рис. 4.2 представлены различные виды активационных функций.

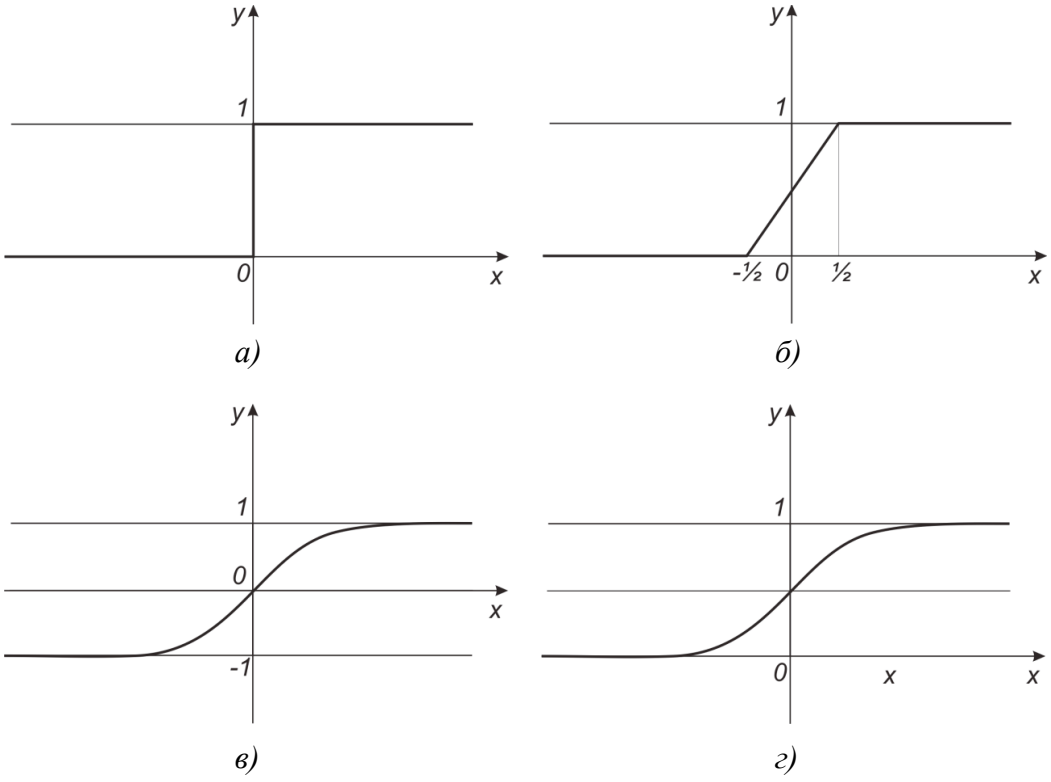


Рис. 4.2. Виды активационных функций: *а* – функция единичного скачка; *б* – линейный порог (гистерезис); *в* – гиперболический тангенс; *г* – сигмоид

Рассмотрим три основных типа функций активации:

1. Функция единичного скачка, или пороговая функция, описывается следующим выражением:

$$f(x)=\begin{cases}1, \text{ если } x \geq 0,\\ 0, \text{ если } x < 0.\end{cases}$$

В литературе она обычно называется функцией Хевисаида. При использовании этой функции выходной сигнал нейрона принимает значение 1, если сигнал на выходе сумматора неотрицательный, и 0 – в противном случае.

2. Кусочно-линейная функция (рис. 4.2, *б*) описывается выражением

$$f(x)=\begin{cases}1, \text{ если } x \geq \dfrac{1}{2},\\ \left|x+\dfrac{1}{2}\right|, \text{ если } \dfrac{1}{2} > x > -\dfrac{1}{2},\\ 0, \text{ если } x \leq -\dfrac{1}{2}.\end{cases}$$

Эту функцию активации можно рассматривать как аппроксимацию нелинейного усилителя.

3. Сигмоидальная функция (т. е. функция S-образного вида). Одна из наиболее распространенных сигмоидальных функций – нелинейная функция с насыщением, так называемая логистическая функция, или сигмоид (рис. 4.2, *г*):

$$f(x) = \frac{1}{1 + e^{-ax}},$$

где a – параметр наклона сигмоидальной функции.

При уменьшении a сигмоид становится более пологим в пределе, при $a = 0$ он вырождается в горизонтальную линию на уровне 0,5. При увеличении a сигмоид приближается по внешнему виду к функции единичного скачка с порогом в точке $x = 0$. Как видно из представленного выражения, выходные значения нейрона лежат в диапазоне (0, 1). Одно из ценных свойств сигмоидальной функции – простое выражение для ее производной:

$$f'(x) = a \cdot f(x) \cdot (1 - f(x)).$$

Следует отметить, что сигмоидальная функция дифференцируема на всей оси абсцисс, что широко используется во многих алгоритмах обучения. Кроме того, она обладает свойством усиливать слабые сигналы лучше, чем сильные, и тем предотвращает насыщение от сильных сигналов, так как они соответствуют областям аргументов, где сигмоид имеет пологий наклон.

Другая широко используемая активационная функция – гиперболический тангенс. В отличие от S-образной функции гиперболический тангенс принимает значения различных знаков, что используется при решении некоторых задач (рис. 4.2, *в*).

4.3. Классификация нейронных сетей

В зависимости от типа входящих сигналов искусственные нейронные сети подразделяются на бинарные и аналоговые. Первые оперируют с двоичными сигналами, и выход каждого нейрона может принимать только два значения: логический ноль (заторможенное состояние) и логическая единица (возбужденное состояние). В аналоговых сетях выходные значения нейронов способны принимать непрерывные значения. Существует еще одна классификация, которая де-

лит нейронные сети на синхронные и асинхронные. В первом случае в каждый момент времени свое состояние меняет лишь один нейрон. Во втором – состояние меняется сразу у целой группы нейронов, как правило, у всего слоя.

Структура нейронных сетей тесно связана с используемыми алгоритмами обучения. В общем случае можно выделить три основных класса нейросетевых структур: однослойные сети прямого распространения, многослойные сети прямого распространения и рекуррентные сети (или сети с обратными связями).

Структура нейронной сети выбирается в соответствии с особенностями и сложностью поставленной задачи. Для решения некоторых типов задач уже существуют оптимальные сегодня конфигурации, описанные в литературе. Если же задача не может быть сведена ни к одному из известных типов, приходится решать сложную проблему синтеза новой конфигурации. При этом следует руководствоваться несколькими основополагающими принципами:

– возможности сети возрастают с увеличением числа ячеек, плотности связей между ними и числом выделенных слоев;

– введение обратных связей наряду с увеличением возможности сети может повлиять на ее динамическую устойчивость;

– повышение сложности алгоритмов функционирования сети (в том числе, например, введение нескольких типов синапсов – возбуждающих, тормозящих и др.) также способствует усилению мощи нейронной сети.

Так как архитектура проектируемой сети сильно зависит от решаемой задачи, дать конкретные рекомендации сложно. В большинстве случаев оптимальный вариант получается на основе интуитивного подбора. Единственное жесткое требование, предъявляемое к структуре сети, – это соответствие размерности вектора входных сигналов сети к числу ее входов.

4.4. Однослойные сети прямого распространения

В многослойной нейронной сети нейроны располагаются по слоям. В простейшем случае в подобной сети существует входной слой узлов источника, информация от которого передается на выходной слой нейронов. Такая сеть называется сетью прямого распространения, или ацикличной сетью. На рис. 4.3 показана простейшая однослойная сеть, состоящая из группы нейронов на входе и трех нейронов на выходе.

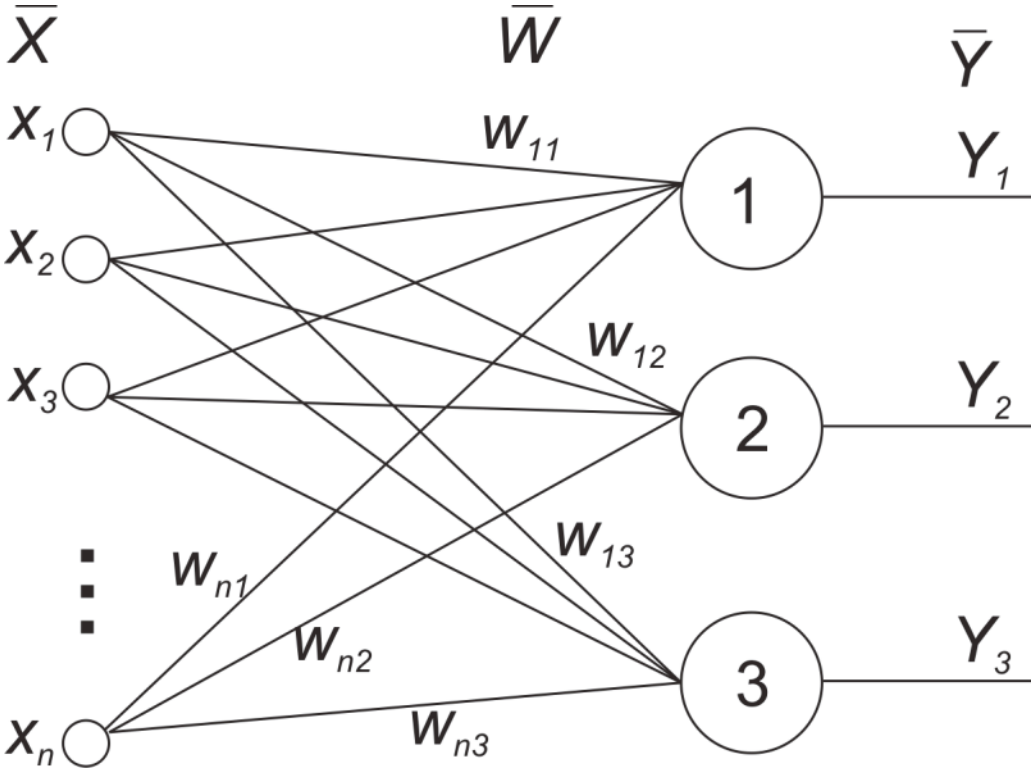


Рис. 4.3. Однослойная нейронная сеть

На n входов поступают некоторые сигналы, проходящие по синапсам на три нейрона, образующих единственный слой этой нейронной сети и выдающих три выходных сигнала:

$$y_j = f\left[\sum_{i=1}^{n} x_i w_{ij}\right],$$

где $j = 1, 2, 3$ – порядковый номер нейрона.

Такая сеть называется однослойной, при этом под единственным слоем подразумевается слой вычислительных элементов. Элементы на входе во внимание не принимаются, так как они не выполняют никаких вычислений.

Весовые коэффициенты синапсов одного слоя нейронов можно свести в матрицу W, в которой каждый элемент w_{ij} задает величину i-й синоптической связи j-го нейрона. Таким образом, процесс, происходящий в нейронной сети, можно записать в виде

$$\overline{Y} = F(\overline{X}\overline{W}),$$

где $\overline{X}$ и $\overline{Y}$ – соответственно векторы входных и выходных сигналов, а $F(\overline{X}\overline{W})$ – активационная функция.

4.5. Многослойные сети прямого распространения

Многослойные нейронные сети характеризуются наличием одного или нескольких скрытых слоев, узлы которых называются скрытыми нейронами. Скрытые нейроны осуществляют связь между входными элементами и выходом сети. Добавляя один или несколько скрытых слоев, можно существенно расширить возможности сети. Такие сети позволяют выделять глобальные свойства данных за счет дополнительных синоптических связей и повышения уровня взаимодействия нейронов.

На рис. 4.4 представлена двухслойная нейронная сеть, полученная из однослойной (см. рис. 4.3) путем добавления второго слоя, состоящего из двух нейронов. Нейроны каждого из слоев сети используют в качестве входных сигналов выходные сигналы предыдущего слоя.

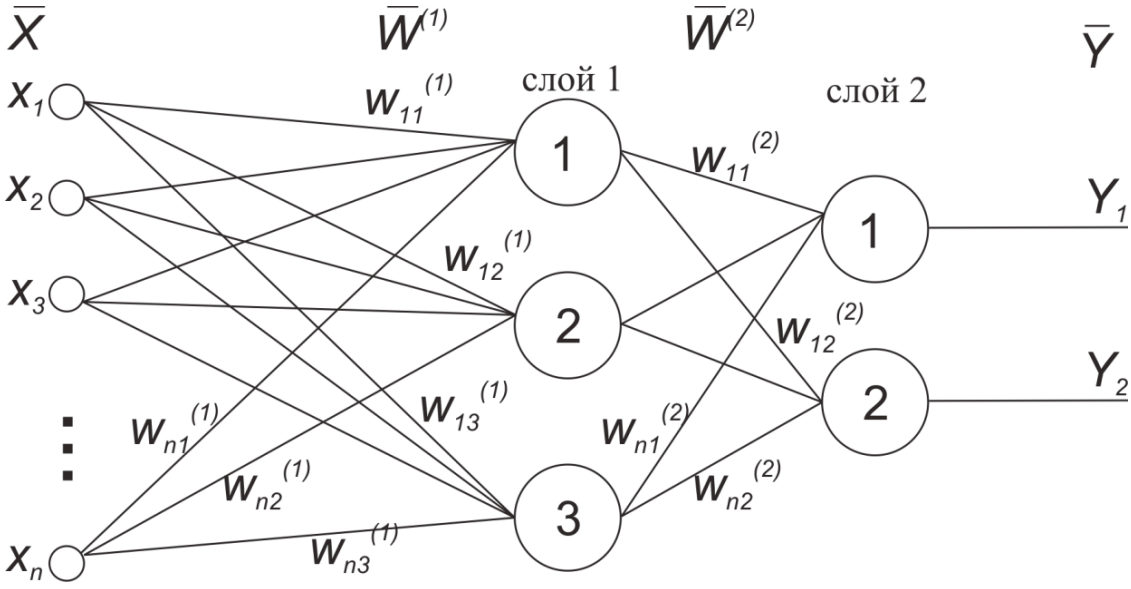


Рис. 4.4. Нейронная сеть с одним скрытым и одним выходным слоями

Для расширения вычислительных возможностей многослойных нейронных сетей можно использовать нелинейные активационные функции. Нелинейность может вводиться и в синоптические связи. В большинстве известных сегодня нейронных сетей для нахождения взвешенной суммы входов нейрона используют формулу

$$d = \sum_{i=1}^{n} x_i w_i \ .$$

Однако в некоторых случаях могут быть и другие зависимости, например:

$$d = \sum_{i=1}^{n} x_i^2 w_i \ .$$

Введение такого рода нелинейности через синапсы увеличивает вычислительную мощь сети, т. е. позволяет, используя меньшее число нейронов с «нелинейными» синапсами, сконструировать нейронную сеть, выполняющую работу обычной нейронной сети с большим числом стандартных нейронов и более сложной конфигурацией.

Как уже говорилось выше, мощность выходного слоя сети, выполняющего окончательную классификацию пространства состояний исследуемого объекта, выбирается исходя из сложности решаемой задачи. Дело в том, что для разделения множества входных образов, например, по двум классам достаточно всего одного выхода. При этом каждый логический уровень «1» или «0» будет обозначать отдельный класс. На двух выходах можно закодировать уже четыре класса и т. д. Однако результаты работы организованной таким образом сети недостаточно надежны. Для повышения достоверности классификации следует ввести избыточность путем выделения каждому классу одного нейрона в выходном слое или, что еще лучше, нескольких, каждый из которых обучается устанавливать принадлежность конкретных состояний входов к определенному классу со своей степенью достоверности, например, высокой, средней и низкой. Такие нейронные сети позволяют проводить классификацию входных неявно выраженных состояний, объединенных в нечеткие (размытые или пересекающиеся) множества. Это свойство позволяет использовать нейронные сети в технической диагностике.

4.6. Сети с обратными связями (рекуррентные)

В рекуррентной нейронной сети, в отличие от сети прямого распространения, имеется по крайней мере одна обратная связь. Например, это может быть сеть из одного слоя нейронов, каждый из которых направляет свой выходной сигнал на входы всех остальных нейронов слоя. На рис. 4.5 показана архитектура рекуррентной сети без скрытых нейронов с обратной связью нейронов с самими собой (ее еще называют сетью Хопфилда).

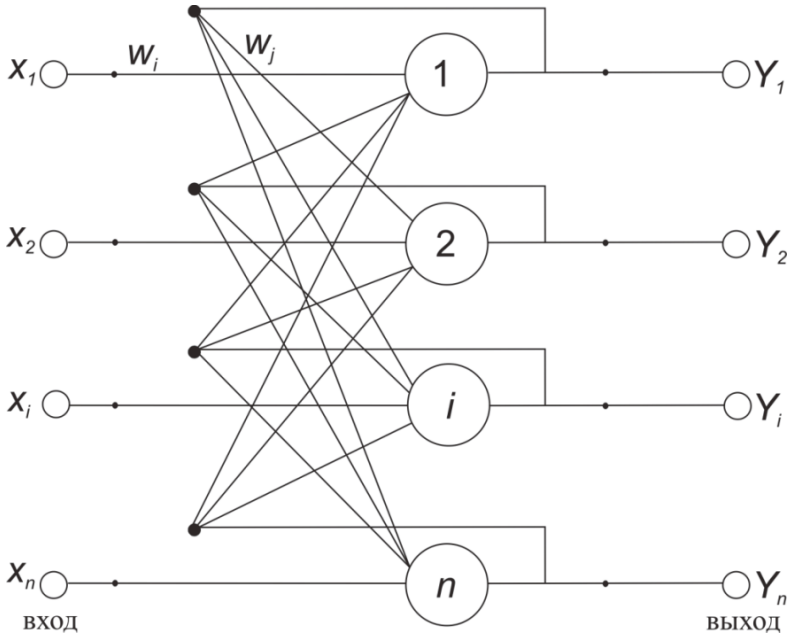


Рис. 4.5. Рекуррентная сеть без скрытых нейронов
с обратной связью нейронов с самими собой

Каждый нейрон связан синапсами со всеми остальными нейронами, а также имеет один входной синапс, через который осуществляется ввод сигнала. Выходные сигналы, как обычно, образуются на аксонах. Рекуррентная сеть может решать задачи распознавания, если есть наборы образцовых данных. Сеть должна уметь из произвольного неидеального сигнала, поданного на ее вход, выделить («вспомнить») соответствующее состояние, т. е. соответствующий образец (если такой есть), или дать заключение о том, что входные данные не соответствуют ни одному из образцов. В общем случае любой сигнал может быть задан вектором. Обозначим вектор, описывающий k-е состояние объекта, через X^k, а его компоненты соответственно через x_i^k. Где $k = 0, 1, \ldots, (m - 1)$, m – число состояний или число запоминаемых образцов. Когда сеть распознает (или «вспоминает») какое-либо состояние на основе предъявленных ей данных, ее выходы будут содержать именно это состояние, т. е. $Y = X^k$-е состояние. В противном случае выходной вектор не совпадет ни с одним образцовым набором данных или состояний. На стадии распознавания весовые коэффициенты синапсов устанавливаются следующим образом:

$$W_{ij} = \begin{cases} \sum_{k=0}^{m-1} X_i^k \cdot X_j^k, \; i \neq j \\ 0, \text{ если } i = j \end{cases},$$

где i и j – индексы соответственно предсинаптического и постсинаптического нейронов; X_i^k, X_j^k – i-й и j-й элементы вектора k-го состояния.

В некоторых случаях сеть не может провести распознавание и выдает на выходе несуществующий образ. Это объясняется ограниченными возможностями сети. Сети Хопфилда имеют число запоминаемых образов т, не превышающее значения равного $0{,}15n$, где n – число нейронов.

Недостаток сетей Хопфилда – их тенденция стабилизироваться в локальном, а не глобальном минимуме. Эта трудность преодолевается с помощью сетей, известных под названием машин Больцмана. В них изменения состояния нейронов обусловлены статистическими, а не детерминированными закономерностями.

4.7. Обучение нейронных сетей

Качество работы нейронной сети в значительной степени зависит от предъявляемого ей в процессе обучения набора данных. Эти данные должны быть типичными для задачи, решению которой обучается сеть. Обучение считается законченным, когда сеть правильно распознает тестовые примеры, а дальнейшее обучение не вызывает значительного изменения весовых коэффициентов. Затем сеть преобразует ранее неизвестные ей данные на основе сформированной в процессе обучения нелинейной модели. Сеть успешно работает, пока существенно не изменится сама модель исследуемого явления. Например, в случае возникновения информации, которая никогда не предъявлялась сети при обучении. В такой ситуации она может быть дообучена с учетом этой новой информации, причем предыдущая информация не теряется, а обобщается с новой.

Качество обучения нейронной сети определяется ее способностью решать поставленные перед ней задачи во время эксплуатации. В теории обучения рассматриваются три основных свойства, связанных с обучением: емкость сети, сложность образцов и вычислительная сложность. Емкость сети обозначает, сколько образцов может запомнить сеть. Сложность образцов – это число обучающих примеров, необходимых для достижения способности сети к распознаванию. Важное значение имеет время, затраченное на обучение. Как правило, время и качество обучения связаны обратно пропорциональной зависимостью и выбираются на основе компромисса.

Существует три принципа обучения: «с учителем», «без учителя» и смешанное (гибридное). Алгоритмы обучения делятся на два боль-

ших класса: детерминированный и стохастический. В первом из них подстройка весов представляет собой жесткую последовательность действий. Во втором она производится на основе действий, подчиняющихся некоторому случайному закону.

Обучение с учителем предполагает, что для каждого входного набора данных существует целевой набор, представляющий собой требуемый выход. Вместе они называются обучающей парой. Обычно сеть обучается на некотором числе таких обучающих пар.

При обучении однослойной сети правильные выходные состояния нейронов заведомо известны и подстройка синаптических связей осуществляется в направлении, минимизирующем ошибку на выходе сети. В многослойных же сетях оптимальные выходные значения нейронов всех слоев, кроме последнего, как правило, неизвестны, и такую нейронную сеть уже невозможно обучить, руководствуясь только величинами ошибок на ее выходах. Проблему можно решить, разработав набор обучающих пар для каждого слоя нейронной сети, что, конечно, не всегда осуществимо. Второй вариант – динамическая подстройка весовых коэффициентов синапсов, в ходе которой выбираются, как правило, наиболее слабые связи, изменяющиеся на малую величину в ту или иную сторону. Но сохраняются только те изменения, которые повлекли уменьшение ошибки на выходе всей сети. Однако такой метод проб, несмотря на кажущуюся простоту, требует большого объема вычислений. Третий вариант – распространение сигналов ошибки от выходов нейронной сети к ее входам в направлении, обратном прямому распространению сигналов в обычном режиме работы. Этот алгоритм обучения получил название процедуры обратного распространения и сегодня является наиболее широко используемым.

Обучение без учителя заключается в подстраивании весовых синапсов. Некоторые алгоритмы предусматривают изменения и структуры сети, т. е. количества нейронов и их взаимосвязей – такие преобразования называются самоорганизацией. На этом принципе построены алгоритмы обучения Хебба.

Сигнальный метод обучения Хебба заключается в изменении весов по следующему правилу:

$$w_{ij}(t) = w_{ij}(t-1) + a \cdot y_i^{(n-1)} \cdot y_j^{(n)},$$

где $y_i^{(n-1)}$– выходное значение i-ого нейрона $(n-1)$ слоя; $y_j^{(n)}$ – выходное значение j-ого нейрона (n) слоя; $w_{ij}(t)$ и $w_{ij}(t-1)$ – весовые коэффициенты синапса, соединяющего эти нейроны на интерациях t и $(t-1)$ соответственно; a – коэффициент скорости обучения.

Существуют также и дифференциальный метод обучения Хебба. При этом методе обучения больше всего обучаются синапсы, соединяющие те нейроны, выходы которых наиболее динамично изменились в сторону увеличения.

Другой алгоритм обучения без учителя – алгоритм Кохонена, предусматривает подстройку синапсов на основании их значений от выходных связей, т. е.:

$$w_{ij}(t) = w_{ij} \cdot (t-1) + a \cdot \left[y_i^{(n-1)} - w_{ij} \cdot (t-1) \right].$$

Он сводит обучение к минимизации разницы между входными сигналами нейрона, поступающими с выходов нейронов предыдущего слоя, и весовыми коэффициентами его синапсов.

Необходимо отметить, что обучение без учителя гораздо более чувствительно к выбору оптимальных параметров, нежели обучение с учителем. Тем не менее с помощью таких сетей можно создать нейронные сети для реально действующих систем с целью оптимизации и прогнозирования технического состояния полиграфического оборудования. Несмотря на некоторые сложности реализации, алгоритм обучения без учителя находит широкое применение. Он используется в наиболее сложных из известных сегодня искусственных нейронных сетей – когнитрон и некогнитрон, максимально приблизившихся в своем воплощении к структуре мозга. Без сомнения, они существенно отличаются от рассмотренных выше сетей и намного более сложны.

Лекция 6. КЛАССИФИКАЦИЯ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ, МЕТОДЫ ИХ ПОЛУЧЕНИЯ И ОСОБЕННОСТИ МОДЕЛИРОВАНИЯ НА РАЗНЫХ УРОВНЯХ

Математические модели классифицируются по характеру отображаемых свойств проектируемого объекта на функциональные и структурные.

Функциональные модели отражают процессы функционирования объекта. Эти модели чаще всего имеют форму систем уравнений. В зависимости от физической природы отображаемых явлений различают модели тепловые, электрические, оптические, гидравлические, кинематические и др.

Структурные модели отражают только структурные (геометрические) свойства объекта. Эти модели могут иметь форму матриц, графов, списков, векторов и выражают взаимное расположение элементов в пространстве. Их используют в случаях, когда задачи структурного синтеза удается ставить и решать, абстрагируясь от особенностей физических процессов в объекте (например, при оформлении конструкторской документации – КД).

На любом иерархическом уровне (микроуровень или макроуровень) математического моделирования можно выделить модели элементов и модели систем. Модель элементов представляет соотношение, связывающее внешние по отношению к этим элементам фазовые переменные.

Полная математическая модель технической системы получается путем объединения модели элементов с общей системой уравнений.

Методы получения математических моделей можно разделить на две группы:

1-я группа – характеризуется использованием в моделях неформальных (эвристических) приемов на этапе выбора математических соотношений (идей) с последующим определением численных значений параметров моделей. Новая идея модели генерируется, например, по методу «мозгового штурма».

В правила метода «мозгового штурма» входят следующие положения: критика не допускается; оценка предложений осуществляется после завершения сессии; чем больше выдвигается идей, тем лучше; чем оригинальнее идея, тем лучше; приветствуются комби-

нации идей; идеи не персонифицируются; численность группы экспертов – 10–15 человек.

Совершенно другим методом генерирования групповой идеи является метод Дельфы. К его характерным чертам относят анонимность, регулируемую обратную связь, групповой ответ. Обработка мнений включает в себя следующие стадии:

1) постановку вопросов в анкетах таким образом, чтобы можно было дать количественную характеристику;

2) проведение опроса в несколько этапов, в ходе которых уточняются ответы и даже вопросы;

3) знакомство всех опрашиваемых с ответами после каждого тура;

4) мотивирование мнения экспертом, если оно отличается от большинства;

5) знакомство с полученной мотивацией всех остальных экспертов;

6) повторение процесса, но при этом участники имеют результаты статистической обработки, а также мнения коллег, несогласованные с мнением большинства. Характер изменения оценок таков, что средняя оценка после интеграций смещается в область, содержащую истинный ответ.

Также в этой группе различают теоретические и экспериментальные методы получения моделей. Теоретические методы основаны на применении физических законов, присущих отображаемым в модели процессам. Основу таких моделей, как правило, составляют системы уравнений, решением которых являются зависимости между фазовыми переменными. Эти модели справедливы в сравнительно широких диапазонах изменения переменных и обычно относятся к алгоритмическим.

Экспериментальные методы основаны на использовании экспериментально полученных зависимостей между параметрами и фазовыми переменными объекта. При этом эксперименты могут проводиться или на самих объектах, или на физических моделях. В процессе преобразования экспериментальных данных в математическую модель производится их аппроксимация, усреднение, статистическая обработка. При этом часто используют методы планирования эксперимента.

2-я группа – методы получения полных математических моделей систем из заданных математических элементов. Методы этой группы инвариантны, т. е. независимы по отношению ко многим областям техники. Например, методы узловых потенциалов могут использоваться в электротехнике, перемещений – в механике и др. Основу

большинства методов 2-й группы составляет один из следующих двух подходов:

а) первый основан на допущении об однонаправленности распространения внешних воздействий от входов к выходам элементов. Этот подход используется при получении моделей логических схем вычислительных устройств, моделировании систем автоматического управления и различных систем массового обслуживания и т. д.;

б) во втором снимается ограничение на однонаправленность модели. Методы второго подхода более сложны в реализации. Их инвариантность обусловлена аналогиями физических систем, поэтому эти методы далее названы методами на основе прямой аналогии.

Лекция 7. ПОСТАНОВКА ЗАДАЧ ОПТИМИЗАЦИИ И ВЫБОР ЦЕЛЕВОЙ ФУНКЦИИ

7.1. Постановка и решение задач оптимизации

На этапе проектирования поиск рационального технического решения осуществляется путем параметрической оптимизации. Методы оптимизации позволяют выбрать наилучший вариант решения технической проблемы из всех возможных при минимуме потерь.

Законы построения технологических машин, в том числе и полиграфических, таковы, что улучшение одного из характерных параметров может приводить к ухудшению других показателей. Буквально на всех этапах проектирования возникают задачи определения степени выигрыша от новых решений и проигрыша, который они повлекут. Например, повышение скорости работы машины без изменения самого принципа технологического процесса, как правило, влечет за собой увеличение габаритов, снижение надежности функционирования машины и качества выполняемых технологических операций, а также повышение напряженности труда оператора (рабочего). Таким образом, оптимизация как выбор варианта решения из некоторого множества подразумевает установление критериев, в соответствии с которыми следует отдать предпочтение одному варианту перед всеми остальными. Выбор критерия – один из важнейших этапов постановки задачи оптимизации, так как все последующие действия направлены на поиск объекта, наиболее близкого к оптимальному по выбранному показателю. В основе построения правил предпочтения лежит целевая функция, количественно выражающая качество объекта и поэтому называемая также функцией качества или критерием оптимальности. Формирование целевой функции всегда выполняется с учетом различных выходных параметров проектируемого изделия. В зависимости от содержательного смысла этих параметров и выбранного способа их сочетания в целевой функции качество объекта будет тем выше, чем больше (максимизация) или чем меньше (минимизация) ее значения.

Выходные параметры, которые могут быть измерены количественно, называются количественными параметрами, а те, которые отображают только качественную сторону объекта, – качественными.

Требуемые соотношения между выходными параметрами и техническими требованиями (ТТ) называют условиями работоспособно-

сти и могут быть записаны в виде: $Y_i < \text{ТТ}_i$, где i лежит в пределах от 1 до k, $\text{ТТ}_j - \Delta Y_j < Y_j < \text{ТТ}_j + \Delta Y_j$, где j лежит в пределах от $k + 1$ до l; ΔY_j – допустимое отклонение j-го выходного параметра ТТ_j от указанного в техническом задании (ТЗ).

Обоснованный вывод о том, насколько удачно то или иное техническое решение, может быть сделан только тогда, когда определены значения всех внутренних параметров, построена математическая модель и выполнены расчеты условий работоспособности.

Внутренние параметры – это независимые переменные, которые полностью и однозначно определяют решаемую задачу проектирования. В качестве таких параметров могут выступать длина звеньев, расстояние, вес, давление, массы, температура и т. д. Число внутренних (проектных) параметров определяет сложность решаемой задачи.

Внутренние параметры, значения которых могут меняться в процессе оптимизации и которые являются аргументами целевой функции, называют управляющими параметрами.

Решение задачи оптимизации можно разбить на два основных этапа:

1. Постановка задачи.
2. Решение задачи, уже имеющей математическую формулировку.

Постановка задачи ведется с учетом назначения реального объекта, целей проектирования и конкретных условий реализации проекта. Эта процедура включает следующие этапы:

– выбор целевой функции и управляющих параметров;

– назначение ограничений;

– нормирование управляющих и выходных параметров и т. п.

Выбор целевой функции. Основная проблема постановки экстремальных задач заключается в формулировке целевой функции. Сложность выбора целевой функции состоит в том, что любой технический объект первоначально имеет векторный характер критериев оптимальности (многокритериальность), причем улучшение одного из выходных параметров часто приводит к ухудшению другого, так как все выходные параметры являются функциями одних и тех же управляемых параметров и не могут изменяться независимо друг от друга. Такие выходные параметры называют конфликтными. При этом целевая функция должна быть одна (принцип однозначности). Сведение многокритериальной задачи к однокритериальной называют сверткой векторного критерия.

В зависимости от того, каким образом выбираются и объединяются выходные параметры, в целевой функции различают частные, ад-

дитивные, мультипликативные, минимаксные статистические критерии оптимальности и т. д.

Назначение ограничений. Ограничения неизбежно появляются при проектировании технических объектов и вытекают из конкретной физической и технологической реализуемости внутренних параметров элементов, ограниченности ресурсов и т. п. При постановке задачи оптимизации учет ограничений иногда бывает принципиально необходим. Так, если целевая функция имеет вид $W(x) = A + B \cdot x$ и не наложены ограничения на параметр x, то задача поиска экстремального значения $W(x)$ становится некорректной. Ограничение суживает область проектных параметров, и искомый экстремум становится условным. Различают прямые и функциональные ограничения. Прямые ограничения имеют вид

$$x_{\text{н}i} \le x_i \le x_{\text{в}i},$$

где $x_{\text{н}i}$ и $x_{\text{в}i}$ – нижнее и верхнее допустимые значения i-го управляемого параметра.

Для многих объектов параметры элементов не могут быть отрицательными, тогда ограничения записываются $x_{\text{н}i} > 0$ (геометрические размеры, электрическое сопротивление, масса и т. д.).

Функциональные ограничения, как правило, представляют собой векторную величину, определяющую условие работоспособности выходных параметров. Функциональные ограничения могут быть двух видов – равенства и неравенства.

Прямые и функциональные ограничения формируют допустимую область поиска. Если функциональные ограничения совпадают с условиями работоспособности, то допустимую область называют областью работоспособности.

Нормирование управляющих и выходных параметров. Пространство управляемых параметров – метрическое. Поэтому при выборе направлений и размеров шагов поиска необходимо вводить ту или иную норму, отождествляемую с расстоянием между двумя точками. Эта норма может быть безразмерной. Возможны различные способы нормирования. Например, способ логарифмического нормирования, достоинством которого является переход от абсолютных приращений параметров к относительным. В этом случае первый управляемый параметр u_i преобразуется в безразмерный x_i:

$$x_i = \ln(u_i / \xi_i),$$

где ξ_i – коэффициент, численно равный единице параметра u_i.

Нормирование выходных параметров можно выполнить с помощью весовых коэффициентов, как в аддитивном критерии.

7.2. Частные критерии оптимальности, примеры их использования, достоинства и недостатки

Частные критерии могут применяться в случаях, когда среди выходных параметров можно выделить один основной параметр $y(X)$, наиболее полно отражающий эффективность проектируемого объекта. Этот параметр и принимают за целевую функцию. Примерами таких параметров являются: для энергетического объекта – мощность, для технологического автомата – производительность, для транспортного средства – грузоподъемность. Для многих технических объектов основным параметром может служить стоимость. Условие работоспособности всех остальных выходных параметров объекта относят при этом к функциональным ограничениям. Оптимизация на основе такой постановки называется оптимизацией по частному критерию.

Достоинство этого подхода – его простота. Существенный недостаток – то, что оптимизация будет производиться только по основному параметру, который принят в качестве целевой функции, а другие выходные параметры вообще не будут оптимизированы.

7.3. Взвешенный аддитивный и мультипликативный критерии, их использование и недостатки

Взвешенный аддитивный критерий применяют тогда, когда можно выделить две группы выходных параметров. В первую группу входят выходные параметры, значения которых в процессе нужно увеличить – $y_j^+(X)$ (например, производительность), во вторую – выходные параметры, значения которых следует уменьшить – $y_j^-(X)$ (например, вес или стоимость). Объединение нескольких выходных параметров, имеющих в общем случае различную физическую размерность, в одной скалярной целевой функции требует их предварительного нормирования. Если принятые параметры можно свести к безразмерным, тогда для минимизации целевой функции свертка векторного критерия будет иметь вид

$$w(X) = \sum_{j=1}^{q} a_j y_j^-(X) - \sum_{j=q+1}^{m} a_j y_j^+(X),$$

где $a_j > 0$ – весовой коэффициент, определяющий степень важности j-го выходного параметра (обычно a_j выбираются проектировщиком и в процессе оптимизации остаются постоянными). Если основные условия работоспособности имеют вид равенств, то целевую функцию, выражающую аддитивный критерий, можно записать в виде

$$w(X) = \sum_{j=1}^{m} a_j \cdot \left[y_j(X) - y_{\mathrm{TT}_j}(X) \right]^2,$$

определяющем среднеквадратичное приближение $y_j(X)$ к значениям заданным техническим требованиям $\mathrm{TT_j}$.

Мультипликативный критерий может применяться, когда отсутствуют условия работоспособности типа равенств и выходные параметры не могут принимать нулевые значения. Тогда минимизируемая мультипликативная целевая функция имеет вид

$$w(X) = \frac{\prod_{j=1}^{q} y_j^-(X)}{\prod_{j=q+1}^{m} y_j^+(X)}.$$

Одним из наиболее существенных недостатков как аддитивного, так и мультипликативного критерия является то, что в постановке задачи не учитываются технические требования, предъявляемые к выходным параметрам.

7.4. Минимаксные (максиминные) критерии, их применение при проектировании

Минимаксные (максиминные) критерии позволяют достичь одной из целей оптимального проектирования – наилучшего удовлетворения условий работоспособности.

Введем количественную оценку степени выполнения j-го условия работоспособности, обозначим ее через Z_j и будем называть запасом работоспособности параметра y_j. Расчет запаса по j-му выходному параметру можно выполнить различными способами, например:

$$Z_j = a_j \left(\frac{\mathrm{TT}_j - y_{j_{\text{ном}}}}{\delta_j} - 1 \right),$$

где a_j – весовой коэффициент; $y_{j\,\text{ном}}$ – номинальное значение j-го выходного параметра; δ_j – величина разброса j-го технического требования; $ТТ_j$ – величина выходного параметра, заданного техническим требованием.

Здесь предполагается, что все соотношения сведены к виду $y_j < ТТ_j$. Если $y_j > ТТ_j$, то необходимо принимать $a_j > 1$ (рекомендуемые значения $5 \le a_j \le 20$). Если желательно достичь выполнения j-го технического требования с заданным допуском, т. е. $y_i = ТТ_j \pm \Delta y_i$; $a_j = 1$, если необходимо получить максимально возможную оценку z_j.

Качество функционирования технической системы характеризуется вектором выходных параметров $Z = (Z_1, Z_2, \ldots, Z_m)$. Поэтому целевую функцию следует формировать как некоторую функцию $\varphi(Z)$ вектора оценок. Если в качестве целевой функции рассматривается запас только того выходного параметра, который в данной точке X является наихудшим с позиций выполнения требований ТЗ, то $W(X) = \min Z_j(X)$, где j изменяется в пределах $1 \le j \le m$ (m – количество запасов работоспособности).

Теперь поставим задачу поиска значения X, которое максимизировало бы минимальный из запасов, т. е. $\max W(X) = \max \min Z_j(X)$, причем это максимальное значение целевой функции должно лежать в допустимой области поиска. Такой критерий оптимизации целевой функции называют максиминным критерием.

Лекция 8–9. МЕТОДЫ БЕЗУСЛОВНОЙ ОПТИМИЗАЦИИ

8.1. Классификация методов поиска экстремума

Решение задач оптимизации в системах автоматизированного проектирования (САПР) ведется с помощью поисковых методов математического программирования, использующих предшествующую информацию для построения улучшенного решения задачи. Большинство постановок задач параметрической оптимизации технических систем сводится к задачам нелинейного программирования, так как целевая функция и ограничения описываются нелинейными зависимостями от вектора управляемых параметров. Если при проектировании удается сформулировать задачу так, что целевая функция и ограничения являются линейными функциями своих аргументов, то имеет место задача линейного программирования. Анализ особенностей постановки задач оптимизации показывает, что задачу параметрического синтеза технических объектов в некоторых случаях можно формулировать как задачу безусловной оптимизации. В зависимости от порядка используемых производных целевой функции по управляемым параметрам методы безусловной оптимизации делят на методы нулевого, первого и второго порядков. Причем наиболее многочисленную группу локальных методов безусловной оптимизации составляют широко применяемые методы нулевого порядка. В методах нулевого порядка информация о производных не используется.

Для методов первого порядка необходимо вычислять как целевую функцию, так и ее первые частные производные. К этим методам относятся метод сопряженных градиентов и метод наискорейшего спуска.

В зависимости от количества управляемых параметров целевой функции различают методы одномерного (метод дихотомии, метод золотого сечения) и многомерного поиска (метод деформируемого многогранника, метод покоординатного спуска, методы случайного поиска). Одномерный поиск может рассматриваться как самостоятельная задача, если аргументом целевой функции является один параметр. Этот же поиск используется в качестве части процедуры многомерной оптимизации в тех случаях, когда необходимо найти оптимальный шаг в выбранном направлении.

Задачу условной оптимизации можно сформулировать как задачу безусловной оптимизации с помощью методов Лагранжа или штрафных функций.

В методах второго порядка организация поиска экстремума ведется с учетом значений целевой функции и ее первых и вторых производных. Методом второго порядка является метод Ньютона.

8.2. Методы одномерного поиска

Обозначим через вектор X искомое значение управляемого параметра, доставляющего локальный экстремум целевой функции $W(X)$. К функции не предъявляются требования дифференцируемости или непрерывности. Предполагается, что для любого X, лежащего в интервале поиска $[a, b]$, значение $W(X)$ может быть вычислено.

Методы одномерного поиска можно разделить на методы последовательного поиска (методы дихотомии, Фибоначчи и золотого сечения) и методы, использующие аппроксимацию функции (методы квадратичной и кубической интерполяции и др.).

Стратегия последовательного поиска X, при которой любая пара вычислений целевой функции $W(X)$ позволяет сузить область поиска (или интервал поиска), осуществляется следующим образом: в интервале поиска вычисляется $W(X)$ в точках X_1 и X_2, лежащих в пределах $a < X_1 < X_2 < b$. Интервал неопределенности можно локализовать путем анализа полученных значений:

если $W(X_1) < W(X_2)$, то X лежит в интервале [a, X_2];

если $W(X_1) = W(X_2)$, то X лежит в интервале [X_1, X_2];

если $W(X_1) > W(X_2)$, то X лежит в интервале [X_1, b].

Стратегия выбора значений X_1 и X_2 для проведения опытов с учетом предыдущих результатов определяет сущность различных методов последовательного поиска.

8.3. Метод дихотомии

Это один из самых простых методов последовательного поиска. Вычисления ведутся по следующему алгоритму. Интервал неопределенности делится на четыре равные части, значения целевой функции вычисляются в трех средних точках (при этом предполагается, что на границах интервала целевая функция известна). Затем выбираются те два отрезка, которые находятся по обе стороны от точки с экстремальным значением целевой функции. Интервал неопределенности

при этом сужается в два раза, при последующих шагах необходимо вычислять значения целевой функции только в двух точках.

В условиях, когда на k-м шаге проводятся два опыта, аргументы X_{1k} и X_{2k} должны выбираться на расстоянии δ/2 справа и слева от середины интервала.

$$X_{1,k} = \frac{a_k + b_k}{2} - \frac{\delta}{2};\ X_{2,k} = \frac{a_k + b_k}{2} + \frac{\delta}{2},$$

где $\delta > 0$ – константа, определяющая расстояние между двумя значениями аргумента; a_k, b_k – границы интервала на k-м шаге.

Вычислив $W(X_{1k})$ и $W(X_{2k})$ и сравнив полученные значения, найдем интервал неопределенности:

если $W(X_{1k}) < W(X_{2k})$, то $a_{k+1} = a_k$, $b_{k+1} = X_{2k}$;

если $W(X_{1k}) = W(X_{2k})$, то $a_{k+1} = X_{1k}$, $b_{k+1} = X_{2k}$;

если $W(X_{1k}) > W(X_{2k})$, то $a_{k+1} = X_{1k}$,, $b_{k+1} = b_k$.

Затем снова вычислим координаты X_1 и X_2 и продолжим поиск.

Коэффициент дробления интервала при этом способе задается зависимостью

$$f = 1/2^{(n-1)/2}.$$

8.4. Метод золотого сечения

В методе золотого сечения целевая функция вычисляется в точках интервала неопределенности, расположенных таким образом, чтобы каждое вычисленное значение целевой функции давало полезную информацию.

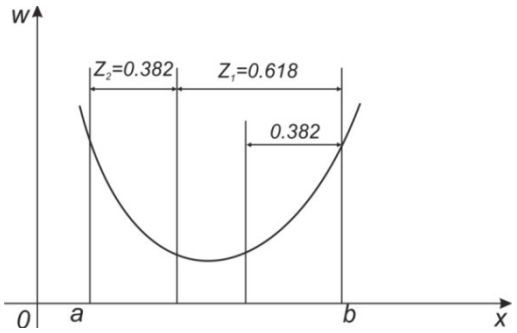


Рис. 8.1. Метод золотого сечения

Сущность метода заключается в следующем: интервал неопределенности делится на две неравные части так, что отношение длины большего отрезка к длине всего интервала равно отношению длины меньшего отрезка к длине большего отрезка, т. е.:

$$\frac{Z_1}{Z} = \frac{Z_2}{Z_1}.$$

Кроме того, $Z_1 + Z_2 = Z$. Из первого уравнения определяем:

$$Z_1^2 = Z \cdot Z_2.$$

Подставив в это выражение значение Z из второго уравнения, получим:

$$Z_1^2 = (Z_1 + Z_2) \cdot Z_2.$$

Разделив обе части на Z_1^2 и решая полученное квадратное уравнение, найдем его корень:

$$Z_1 / Z_2 = 0{,}618\ .$$

При этом методе на первом шаге нужно вычислить два значения целевой функции в точках X_1 и X_2, затем выбрать ту часть интервала, где значение целевой функции W ближе к экстремуму. Значения X_1 и X_2 определяются следующим образом:

$$X_1 = a_k + 0{,}382\,(b - a),\ X_2 = b_k - 0{,}382\,(b - a).$$

На втором и всех последующих шагах потребуется вычислить уже только одно значение целевой функции. На каждом шаге расчетов интервал неопределенности сокращается в 1,618 раза, а коэффициент $f = 0{,}618^{n-1}$.

8.5. Особенности поиска при максиминных постановках задач оптимизации

Максиминный метод оптимизации применяется для решения задач условной оптимизации.

Максиминная постановка задач параметрической оптимизации – одна из наиболее перспективных при проектировании технических объектов. Это объясняется тем, что в результате оптимизации улучшаются запасы работоспособности практически для всех выходных параметров проектируемого изделия. Однако решение этих задач при такой постановке имеет специфику. Прежде всего, это касается условий формирования целевой функции, которая называется функцией с максиминным критерием и выглядит $W(X) = \min Z_j(X)$, где $Z_j(X)$ – оценка степени выполнения j-го условия работоспособности, а j – изменяется в пределах $1 < j < m$ (m – количество запасов работоспособности). Эта оценка в процессе поиска формируется следующим образом.

В каждой из отображающих точек в диапазоне поиска находятся выходные параметры оптимизируемого объекта и по ним определяются оценки Z_j.

Среди всех оценок в точке X_k находят минимальную, пусть, например, это будет оценка Z_p. В этом случае целевая функция $W(X) = Z_p(X)$. До момента смены оценки Z_p на другую гиперповерхность целевой функции является гладкой, что позволяет провести максимизацию этой целевой функции одним из методов нелинейного программирования. Происходит улучшение оценки Z_p. Однако в силу конфликтности выходных параметров улучшение оценки Z_p приводит к ухудшению других оценок. Например, увеличение прочности элемента конструкции приводит к увеличению ее массы и стоимости. Поэтому на определенном этапе поиска оказывается, что минимальной становится другая оценка, например Z_q. Такая смена минимальных оценок может наблюдаться для разных условий.

На представленном рис. 8.2 показан случай, когда на $(k + 1)$-м шаге поиска в качестве целевой функции будет фигурировать функция $W(X) = Z_q(X)$.

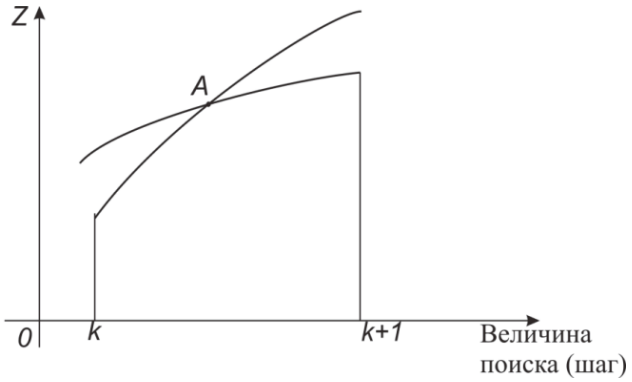


Рис. 8.2. Максиминный метод параметрической оптимизации

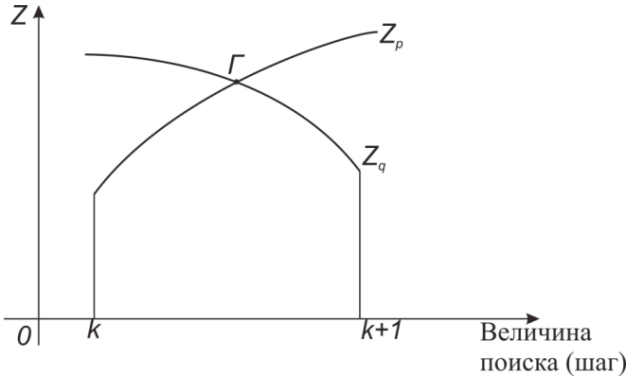


Рис. 8.3. Анализ поведения оценок

Как видно из рис. 8.3, оценки Z_p и Z_q имеют тенденцию к увеличению, и смену целевой функции можно трактовать как переход с одной гиперповерхности поиска на другую. Анализ поведения оценок, отраженный на рис. 7.3, показывает, что ограничиться простой сменой оценок в целевой функции нельзя. Искомый максимум должен находиться на гиперповерхности пересечений двух гиперповерхностей $Z_p(X)$ и $Z_q(X)$. Эта гиперповерхность пересечений является гребнем гиперповерхности целевой функции, а точка пересечения оценок $Z_p(X)$ и $Z_q(X)$ является точкой гребня. В этом случае безусловный поиск целесообразно заменить на условный: вводится ограничение типа равенства $Z_p(X) = Z_q(X)$, а в качестве целевой функции можно рассматривать прежнюю оценку.

Анализ особенностей формирования целевой функции показывает, что функция минимума $W(X)$ не может быть гладкой. В точках гребней целевая функция не дифференцируема, поэтому один из наиболее эффективных алгоритмов решения задач в максиминной постановке – метод проекции градиента. Использование этого метода возможно потому, что уравнения гребней удается сформулировать в виде равенств конфликтных оценок $Z_p(X) = Z_q(X)$ и рассматривать их как ограничения задачи.

8.6. Методы случайного поиска

Методы случайного поиска классифицируются как методы многомерного поиска. Идея заключается в том, чтобы перебором случайных значений управляемых параметров найти экстремум значения целевой функции. В этих методах направление поиска выбирается случайным образом – на основании генерации ЭВМ случайных чисел посредством генератора случайных чисел (ГСЧ). Среди многих разновидностей методов случайного поиска простейшим является метод Монте-Карло. На $(k + 1)$-м шаге поиска определяется случайная точка X_{k+1}, вычисляется значение $W(X_{k+1})$ и сравнивается со значением, полученным на предыдущем шаге. Если $W(X_{k+1}) < W(X_k)$, то запоминаются координаты точки и новые значения целевой функции и поиск экстремума продолжается в этом направлении. Сущность метода оптимизации определяется двумя условиями: направлением поиска и шагом поиска. Исходя из указанных условий, прекращение поиска зависит от величины выборки случайных чисел.

Теоретически при достаточно большом количестве выборок случайных чисел можно достичь высокой точности определения экстре-

мума. Однако приемлемая точность вычислений требует больших вычислительных затрат. Например, если экстремум определяется с точностью ε, то при выборе случайных точек необходимо хотя бы один раз попасть в ε – окрестность точки экстремума.

Если производится k выборок случайных точек, то вероятность попадания, хотя бы одной из них в окрестность точки экстремума ε составляет $P(k) = 1 - (1 - \varepsilon^n)^k$. Следовательно, величина выборки k случайных точек, необходимая для того, чтобы с вероятностью P можно было бы утверждать, что найденное с точностью ε оптимальное значение соответствует истинному, равна

$$k \approx (1/\varepsilon)^n \ln\left[1/(1-P)\right],$$

где ε – окрестность точки экстремума; n – размерность пространства (Евклидова, $n = 2$); P – вероятность попадания случайной точки в ε-окрестность.

Например, для двумерной задачи ($n = 2$) при $P = 1/2$ и $\varepsilon = 10^{-3}$ нужно вычислить выборку случайных точек:

$$k = (1/10^{-3})^2 \cdot \ln\left[1/(1-1/2)\right] = 10^6 \ln 2 = 0{,}69 \cdot 10^6.$$

Поэтому в используемых на практике методах случайного поиска вычисления производятся с использованием алгоритма численной оптимизации, т. е. попытка достичь успеха реализуется путем выбора либо новой точки поиска X_{k+1}, либо случайного нового направления. Использование метода Монте-Карло поясним следующим примером.

Предположим, нужно собрать компьютер, в состав которого входят три важные детали. Половина деталей закуплена на Тайване, а половина – в материковом Китае. Изделие считается бракованным, если в него попадут все три детали, изготовленные в Китае. Выбор деталей со склада имеет равномерное распределение (рис. 8.4).

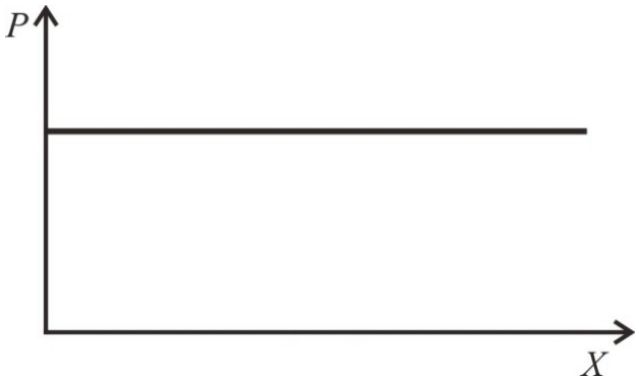


Рис. 8.4. Равномерное распределение вероятности

Предположим, у нас имеется монета и при выпадении «орла» будем считать, что элемент выбран из Тайваня, а «решки» – из Китая. Если одновременно бросить три монеты и при этом выпадут три «решки», то компьютер считается неисправным. Таким образом, мы сымитировали ситуацию и разыграли случайное число, которое имеет равномерное распределение. Вероятность первого события (элемент из Тайваня) равно 0,5, и вероятность второго события (элемент из Китая) тоже равна 0,5.

Разыгрывать случайные числа таким способом крайне неудобно, поэтому разработаны алгоритмы генерации случайных чисел, в которых случайные цифры имитируют равномерное распределение случайных величин. Воспользуемся вышеуказанным алгоритмом генерации случайных чисел и внесем результаты генерации в табл. 8.1. Примем также, что при вероятности меньшей или равной 0,5 элемент будет отнесен к Тайваню, в противном случае – к Китаю.

Таблица 8.1

Номер реализации	1-я деталь	2-я деталь	3-я деталь	Результат
1	0,86	0,51	0,56	Брак
2	0,91	0,86	0,41	Годен
3	0,68	0,68	0,65	Брак
4	0,22	0,72	0,58	Годен
5	0,75	0,24	0,52	Годен
6	0,76	0,77	0,30	Годен
7	0,48	0,25	0,87	Годен
8	0,87	0,11	0,38	Годен
9	0,38	0,47	0,54	Годен
10	0,90	0,79	0,51	Брак

Теперь разделим количество случаев с браком на общее количество реализации. Получим 3/10 = 0,3. Данную задачу можно решить и аналитически. В этом случае ответ будет 0,125. Если продолжим дальнейшие реализации, то можем получить ответ, близкий к расчетному. Какой вывод можно сделать из этого примера? Очевидно, что результат в данном методе напрямую связан с количеством реализаций. И ошибка метода пропорциональна величине $1/\sqrt{N}$. Следовательно, сходимость к истинному результату требует значительного числа реализаций, а значит, необходим «хороший» генератор случайных чисел.

Метод Монте-Карло часто используется в тех случаях, когда другие методы применить невозможно или нежелательно, например, для оценки надежности сложных систем.

Допустим, требуется оценить возможность безотказной работы полиграфической системы допечатной обработки цифровой информации, состоящей из трех аппаратно-программных узлов.

Узел 1 состоит из двух элементов А и Б и выходит из строя при отказе одновременно обоих элементов А и Б, т. е. если только элемент А или только элемент Б неисправен, то узел исправен. Предположим, что вероятность безотказной работы этих элементов за определенный промежуток времени составляет соответственно 0,8 и 0,9.

Узел 2 неисправен при отказе элемента В, и вероятность безотказной работы за тот же промежуток времени равна 0,8.

Узел 3 состоит из трех элементов Г, Д, Е и выходит из строя, если неисправен любой из этих элементов. Вероятность их безотказной работы равна соответственно 0,7, 0,8, 0,9.

Неисправность любого узла вызывает неисправность всей системы.

Для оценки работы каждого узла используется также таблица случайных величин.

Для запуска испытаний необходимо генерировать случайные числа и сравнивать их величину с пороговым значением, при котором происходит отказ элемента. Если выбранное значение превосходит пороговое, то отмечают отказ, в противном случае элемент исправен. После определения отказа либо исправности элемента определяют состояние узла, а затем системы. Результат испытаний приведен в табл. 8.2.

Разделив количество исправных состояний на общее количество реализации, получим, что вероятность исправного состояния такой полиграфической системы составляет $4/11 = 0,36$.

Следовательно, многократная реализация вышеописанной модели и усреднение результатов позволят получить необходимый результат и дать оценку бесперебойной работы полиграфической системы допечатной обработки цифровой информации.

Таблица 8.2

Номер реализа-ции	Реализация		Результат для узла 1	Реали-зация	Результат для узла 2	Реализация			Результат для узла 3	Общий результат
	А	Б		В		Г	Д	Е		
1	0,66	0,15	И	0,54	И	0,25	0,02	0,85	И	И
2	0,18	0,18	И	0,80	Н	0,59	0,89	0,68	Н	Н
3	0,83	0,33	И	0,46	И	0,27	0,25	0,68	И	И
4	0,03	0,17	И	0,45	И	0,86	0,33	0,81	Н	Н
5	0,32	0,66	И	0,43	И	0,49	0,92	0,92	Н	Н
6	0,38	0,96	И	0,78	И	0,93	0,42	0,42	Н	Н
7	0,41	0,68	И	0,39	И	0,35	0,66	0,66	И	И
8	0,75	0,12	И	0,02	И	0,45	0,20	0,20	И	И
9	0,29	0,45	И	0,65	И	0,58	0,88	0,88	Н	Н
10	0,53	0,31	И	0,81	Н	0,67	0,24	0,24	И	Н
11	0,75	0,89	И	0,79	И	0,13	0,02	0,57	И	И

8.7. Схема использования метода Монте-Карло при исследовании систем со случайными параметрами

Построив модель системы массового обслуживания, на ее вход подают входные сигналы от генератора случайных чисел (ГСЧ). ГСЧ устроен так, что он выдает равномерно распределенные случайные числа из интервала от 0 до 1. Так как некоторые события более вероятны, а другие – менее, то случайные числа от генератора подают на датчик (ДСЧ), который преобразует их в заданный закон распределения вероятности. Далее модель обрабатывает входной сигнал и получает выходной, который является случайным событием. Если процедуру повторять многократно, то на выходе появится массив случайных чисел, который формируется в блоке накопления статистики (НС). Затем исследуют вероятностное распределение выходного сигнала в блоке расчета статистических показателей (РСП) и делают заключение о свойствах объекта, который моделируют. В блоке НС анализируют степень достоверности результата и определяют необходимое количество статистических испытаний. При малом числе испытаний результат может оказаться недостоверным.

Рис. 8.5. Схема использования метода Монте-Карло

Например, событие A совершилось в результате проведенных 200 экспериментов 50 раз. Согласно методу Монте-Карло, это означает, что вероятность совершения события

$$P_A = 50 / 200 = 0{,}25.$$

Вероятность того, что событие не совершится, равна соответственно 1 – 0,25 = 0,75. Оценка вероятностей производится по следующей формуле:

$$P = n / N.$$

Лекция 10. ДИНАМИЧЕСКАЯ МОДЕЛЬ И ЕЕ ХАРАКТЕРИСТИКА

В общем случае, если выходные переменные параметры задаются как функции времени t, а в уравнения, описывающие моделируемый процесс или объект, входят производные по времени (или интегралы) фазовых переменных, математическая модель такого объекта представляет систему дифференциальных или интегральных уравнений и называется динамической моделью.

Исследования динамики механизма, прежде всего, предполагает исследование движения его массы при воздействии на нее сил. Укрупненно действующие силы можно подразделять на инерционные, внутреннего сопротивления, упругости, внешние (технологические, сопротивления) и др.

Реальные тела имеют бесконечное число степеней свободы. Для решения же практических задач необходимо разумно ограничить их количество так, чтобы это не искажало закономерности движения масс. Исходя из этого, при составлении динамической модели следует:

– выделить наиболее важные степени свободы;

– отобразить инерционные свойства системы массами или моментами инерции, которые необходимо сосредоточить в определенных точках;

– соединить эти точки кинематическими связями или учесть взаимодействие между ними.

Практически эти шаги сводятся к тому, что выделяются наиболее массивные элементы, а также наиболее податливые (наименее жесткие), и эти элементы связываются в кинематические цепи.

Одномассовая динамическая модель описывается уравнением

$$m\frac{d^2q}{dt^2}+\mu\frac{dq}{dt}+cq=\sum_{j=1}^{n}F_j\,,$$

где первое слагаемое – инерционные силы; второе слагаемое – силы внутреннего сопротивления; третье слагаемое – силы упругости; F_j – внешние силы; n – число этих сил.

В других случаях, особенно при моделировании цикловых механизмов, предпочтительно использовать модели, отражающие моменты, в которых в начале обобщенных координат выбираются углы поворота. При этом инерционные характеристики выражаются моментами инерции звеньев (J_0, J_1, J_2), учитывающими упруго дисси-

пативные воздействие (ψ и c), а также кинематические связи, отображаемые функциями положений Π_n.

Для математического описания динамической модели со сосредоточенными параметрами наиболее часто используют уравнения Лагранжа. Для этого необходимо:

– выбрать обобщенные координаты по числу степеней свободы модели;

– определить точки сосредоточения масс или моментов инерции;

– привести к этим точкам массы и силы инерции;

– привести упругодиссипативные связи в механизме к безынерционным связям между инерционными элементами системы;

– определить число и направление обобщенных координат;

– составить уравнения Лагранжа в вышеупомянутых координатах с учетом связей, ограничивающих функционирование механизма;

– представить уравнения Лагранжа в виде, возможном и удобном для математического разрешения.

С этой целью систему надо представить в виде либо голономной, либо неголономной и в зависимости от этого использовать тот или иной математический аппарат.

Если уравнения связи не содержат производных от координат или они интегрируемы, то они называются голономными. Примером голономной связи является функция положения

$$\Pi(\varphi) = 0\,,\ \Pi(\varphi_1, \varphi_2, ..., \varphi_n) = 0\,.$$

Приведем следующий пример составления динамической модели.

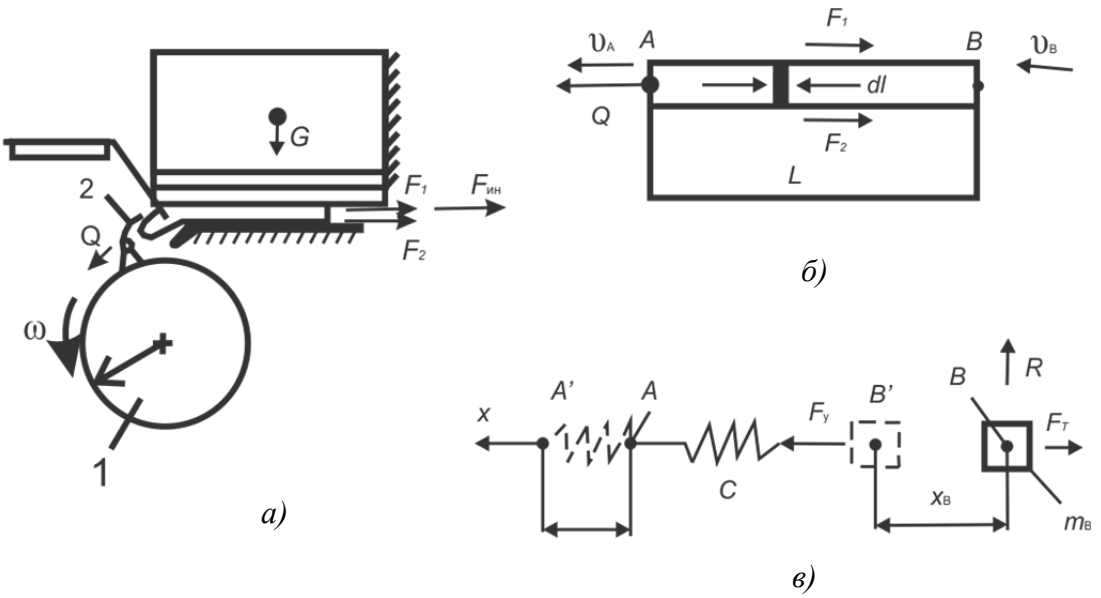
Специфическая особенность функционирования механизмов полиграфических машин – значительное влияние на них упругости бумажного полуфабриката.

Рассмотрим процесс вывода тетради из магазина тетрадного самонаклада, технологическая схема которого представлена на рисунке, *а*.

Выводной цилиндр *1* вращается с постоянной скоростью ω = const. Клапан *2* захватывает корешок тетради и вытаскивает ее из-под стопы. При этом развивается тянущее усилие:

$$Q = \left\{F_1 + F_2\right\} e^{\gamma f}\ F_{\text{ин}}\,,$$

где F_1 – сила трения между стопой и тетрадью; F_2 – сила трения между дном магазина и тетрадью; γ – угол огибания гибким телом твердой поверхности; f – коэффициент трения; $F_{\text{ин}}$ – инерционная сила, которая развивается в момент захвата тетради.



Технологическая схема процесса вывода тетради из магазина (*а*) и его эквивалентная (*б*) и динамическая (*в*) модели; *1* – выводной цилиндр; *2* – клапан

Скорость тетради практически мгновенно возрастает от 0 до окружной скорости цилиндра $\upsilon_ц$. Оказалось, что $F_{ин}$, возникающая при этом, примерно на два порядка превышает силы трения. Проанализируем это явление как разгон упругого тела массой m, обладающего жесткостью C.

На рисунке представлены эквивалентная и динамическая модели этого процесса. При разгоне упругого тела отдельные поперечные сечения тетради *dl* (рисунке, *б*) будут развивать разные скорости. Возьмем, например, сечение A, в котором происходит фиксация тетради клапаном. Скорость этого элемента тетради будет равна $\upsilon_A = \omega \cdot r$. А остальные элементы будут совершать сложное движение – переносное со скоростью υ_A и относительное – колебательное по отношению к сечению A. Такая система обладает бесконечным числом степеней свободы. Поэтому сделаем допущение, приводя массу тетради в точку B, что колебания происходят только в горизонтальном направлении. Такая система моделируется так, как показано на рисунке, *в*.

Уравнение движения массы $m_в$ в точке B можно представить в виде

$$m_в \ddot{x} = F_у - F_т \ \text{sign}\, \dot{x}, \qquad (10.1)$$

где $F_у$ – сила упругости; $F_т$ – сумма сил трения; sign $\dot{x}$ – знак скорости $\dot{x}$.

Точка A начнет перемещаться в момент захвата, а точка B в этот момент может еще оставаться на месте.

Движение $m_в$ начнется после преодоления силы трения покоя $F_{тп}$. Поэтому $x_0 = 0,\ \dot{x} = 0$ в начальный момент при $t = 0$.

Время t отсчитываем от момента, когда груз приходит в движение. К этому моменту времени пружина растягивается на величину δ, определяемую условием

$$c \cdot \delta = F_{тр}\,. \tag{10.2}$$

Рассмотрим движение груза при условии $\dot{x} \geq 0$. В этом случае дифференциальное уравнение движения будет

$$m\ddot{x} = c(vt - x + \delta) - F_{тр}\,. \tag{10.3}$$

С учетом (10.1) и после деления на m получим

$$\ddot{x} = k^2 x = k^2 vt\,, \tag{10.4}$$

где

$$k = \sqrt{\frac{c}{m}}\,.$$

Общее решение уравнения (10.4)

$$x = A\cos kt + B\sin kt + vt\,. \tag{10.5}$$

Используя начальные условия, находим

$$A = 0,\ B = \frac{v}{k}, \tag{10.6}$$

так что искомое частное решение

$$x = vt - \frac{v}{k}\sin kt\,. \tag{10.7}$$

Скорость груза

$$v = \omega r\,,$$

$$\dot{x} = v - v\cos kt = v(1 - \cos kt) = 2v\sin^2\frac{kt}{2} \geq 0\,, \tag{10.8}$$

т. е. груз в любой момент времени имеет неотрицательную скорость.

Усилие, действующее на тетрадь, определяется упругой силой:

$$F_{упр} = c\delta = c(vt - x + \delta) = F_{тр} + c\frac{v}{k}\sin kt = F_{тр} + \sqrt{mc}v\sin kt\,. \tag{10.9}$$

Максимальное значение усилия

$$F_{\max} = F_{\text{тр}} + c\frac{v}{k} = F_{\text{тр}} v\sqrt{mc} \,. \qquad (10.10)$$

Если $F_{\text{тр}} \ll F_{\max}$, то $F_{\max} \sim \nu, \sqrt{m}, \sqrt{c}$.

Наличие сил трения, пропорциональных скорости движения, приведет к затухающим колебаниям относительно скорости (v) захвата. Вместо полученного соотношения (8.8), зная величину задержки τ, по времени начала движения груза можно оценить эффективную жесткость тетради:

$$\delta = v\tau;\; c = \frac{F_{\text{тр}}}{v\tau}\,.$$

Лекция 11–12. КОДИРОВАНИЕ ИЗОБРАЗИТЕЛЬНОЙ ИНФОРМАЦИИ

11.1. Дискретизация и квантование изображений

При рассмотрении понятия энтропии непрерывных сигналов необходимо прибегать к их дискретизации во времени и квантованию по амплитуде (уровню). В первом случае интервал дискретизации должен выбираться в соответствии с теоремой Котельникова.

Теорема Котельникова. Произвольный сигнал с ограниченным спектром, не содержащим частот выше f_b, может быть полностью восстановлен, если известны его дискретные значения, взятые через промежутки времени:

$$T_{\text{И}} = \frac{1}{2f_b}.$$

Этот сигнал восстанавливается с помощью следующего ряда:

$$x(t) = \sum_{k=-\infty}^{\infty} \frac{\sin(\omega_b t - k\pi)}{\omega_b t - k\pi}.$$

Для точного восстановления исходного изображения, согласно теореме Котельникова, необходимо выполнение следующих условий:

– спектр оптических изображений должен быть ограничен частотами $\omega_{\beta max}$ и $\omega_{\alpha max}$;

– шаг дискретизации должен приниматься не меньшим

$$\alpha = \frac{\pi}{\omega_{\alpha\,max}},\ \beta = \frac{\pi}{\omega_{\beta\,max}}.$$

При дискретизации изображения нарушение теоремы Котельникова ведет к его искажению.

В результате дискретизации непрерывное оптическое изображение представляется в виде совокупности дискретных элементов – пикселей, имеющих обычно квадратную форму. При переходе к дискретному квантованному изображению важно оценить статистическую связь между оптическими плотностями соседних пикселей по строке и по кадру. Совокупность коэффициентов корреляции r_c между различно отстоящими друг от друга пикселями можно представить в виде корреляционной матрицы. Она для строки имеет вид:

$$R_C = D_X \begin{vmatrix} 1 & r_C & r_C^{\,2} & r_C^{\,3} & \dots & r_C^{\,n} \\ r_C & 1 & r_C & r_C^{\,2} & \dots & r_C^{\,n-1} \\ r_C^{\,2} & r_C & 1 & r_C & \dots & r_C^{\,n-2} \\ r_C^{\,3} & r_C^{\,2} & r_C & 1 & \dots & r_C^{\,n-3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_C^{\,n} & r_C^{\,n-1} & r_C^{\,n-2} & r_C^{\,n-3} & \dots & 1 \end{vmatrix} \qquad (11.1)$$

Аналогичный вид имеет и матрица переходов от одного пикселя к другому по кадру:

$$R_K = D_X \begin{vmatrix} 1 & r_K & r_K^{\,2} & r_K^{\,3} & \dots & r_K^{\,n} \\ r_K & 1 & r_K & r_K^{\,2} & \dots & r_K^{\,n-1} \\ r_K^{\,2} & r_K & 1 & r_K & \dots & r_K^{\,n-2} \\ r_K^{\,3} & r_K^{\,2} & r_K & 1 & \dots & r_K^{\,n-3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_K^{\,n} & r_K^{\,n-1} & r_K^{\,n-2} & r_K^{\,n-3} & \dots & 1 \end{vmatrix} \qquad (11.2)$$

Матрицы переходов от одного пикселя к другому характеризуют собой дискретный Марковский процесс перехода от оптической плотности одного пикселя к оптической плотности другого. Этот дискретный процесс является аналогом одномерной плотности распределения:

$$F(x) = D_X e^{-\alpha|x|}.$$

Корреляционную матрицу изображения можно представить в виде произведения матриц переходов по строкам и по кадру:

$$R = R_C R_K .$$

Коэффициент корреляции между пикселями по диагонали в этом случае будет равен произведению $r_D = r_c r_K$. Непосредственные расчеты при обработке дискретной информации показывают, что представление матриц $R_с$ и $R_к$ в виде (11.1) и (11.2) является достаточно точным при расстоянии между элементами изображения до 5 пикселей.

11.2. Общие принципы кодирования изображений

При квантовании и дискретизации изображений в ЭВМ вводится первичное кодовое описание. Количество информации в этом описании определяется размерами изображения, линиатурой (размерами пикселей) и числом градаций оптической плотно-

сти. Первичное описание изображений связано с использованием растровых форматов изображения. Под **форматом** понимается размещение данных изображений на машинном носителе, в котором зафиксированы координаты каждого пикселя и его оптическая плотность. На размер файла в этом случае влияют число пикселей в изображении и число разрядов, отводимых на один пиксель для определения его цветности. При описании цветных изображений иногда используется до 24 и 32 разрядов на один пиксель. Это означает, что изображение размером 1024×768 пикселей требует объемов файлов до 3,0 Мб. Даже файлы черно-белых изображений при том же числе пикселей и числе градаций оптической плотности, равной 256 и требующей 8 разрядов для описания каждого пикселя, должны иметь объем 768 Кб.

Поэтому используется обработка изобразительной информации с переходом ко вторичному кодовому описанию изображения. Вторичное описание предусматривает применение других форматов представления изображений, которые требуют значительно меньшего объема файлов по сравнению с растровыми форматами. Для этого используются программы-трансляторы форматов.

Требования к первичному и вторичному кодовому описанию различны. Первичное описание дает возможность восстанавливать исходное изображение с требуемой точностью. Вторичное кодовое описание должно существенным образом «сжать» информацию без потери ее качества. Уменьшение объема информации об изображении называется сжатием визуальных данных. В этом случае данные представляются в виде вектора признаков или символического описания, позволяющего распознавать элементы изображения. Сжатие информации не только экономит объем информации, но и обеспечивает эффективное восстановление изображения. Принцип сжатия состоит в том, что изображение описывается с помощью коэффициентов, наименее коррелированных между собой. Учет такой корреляции может быть осуществлен в дальнейшем при восстановлении изображения.

В настоящее время используется большое количество форматов представления изобразительной информации, обеспечивающих сжатие данных. В них применяются разнообразные методы сжатия данных.

Наиболее простой метод сжатия данных – учет неравновероятности градации оптической плотности в изображении. Знакоизбыточность, вносимая при пренебрежении неравновероятности оптических плотностей пикселей, невелика, и сократить объем памяти ЭВМ можно не более, чем в 1,5 раза.

Более существенного сокращения объема требуемой памяти ЭВМ можно достигнуть с учетом корреляции между значениями оптической плотности соседних пикселей. Если ввести в рассмотрение условную вероятность $P(i|j)$ того, что отсчет оптической плотности имеет значение i при условии наличия отсчета j в соседнем пикселе, то условная энтропия будет

$$H_{x_1}(x_2) = -\sum_{i=0}^{m}\sum_{j=0}^{m} P(i,j)\log_2 P(i \mid j),$$

где $P(i,j) = P(j)P(i|j)$ – вероятность того, что отсчеты оптической плотности в соседних пикселях равны i и j.

В соответствии с существующей зависимостью между совместной и условной энтропиями двух независимых событий энтропию двух независимых событий x_1 и x_2 можно выразить

$$H(x_1, x_2) = H(x_1) + H_{x_1}(x_2). \qquad (11.3)$$

В том случае, когда отсчеты оптической плотности в точках i и j независимы, энтропия двух пикселей будет

$$H(x_1, x_2) = H(x_1) + H(x_2). \qquad (11.4)$$

Исходя из свойств энтропии взаимосвязанных событий, найденная по формуле (11.3) энтропия меньше энтропии, найденной по формуле (11.4). А это говорит об уменьшении количества информации в отсчетах с коррелированными значениями оптической плотности.

Важнейшая характеристика метода сжатия визуальных данных – точность восстановления изображения. Если обозначить $f(i,j)$ – первичное кодовое описание изображения, а $g(i,j)$ – вторичное описание после сжатия данных, то дисперсия отклонения функции $g(i,j)$ от функции $f(i,j)$ будет

$$D_x = \frac{1}{m^2}\sum_{i=0}^{m}\sum_{j=0}^{m}[g(i,j) - f(i,j)]^2. \qquad (11.5)$$

Это отклонение является помехой, поэтому отношение сигнал/помеха выходного изображения определяется

$$k = \frac{\sum_{i=0}^{m}\sum_{j=0}^{m}[g(i,j)]^2}{\sum_{i=0}^{m}\sum_{j=0}^{m}[g(i,j) - f(i,j)]^2}. \qquad (11.6)$$

Замечание. В некоторых случаях объективные оценки методов сжатия данных, вычисленные по формулам (11.5) и (11.6), не всегда отражают реальное визуальное качество изображения. Это объясняется особенностями зрительной системы человека. Поэтому окончательное решение о качестве сжатия данных принимается субъективным методом путем визуальной оценки изображения.

11.3. Методы кодирования тоновых изображений

11.3.1. Эффективное кодирование

Эффективное кодирование тоновых изображений предусматривает учет неравномерного распределения оптической плотности. При этом сжатие данных достигается за счет того, что для представления наиболее вероятных оптических плотностей используются короткие кодовые слова, а для наименее вероятных – длинные. Примерами таких преобразований информации являются алгоритм Шеннона-Фано и алгоритм Хаффмена, которые будут рассмотрены далее.

Особенность этого метода в том, что необходимо учитывать статистическую зависимость между оптическими плотностями соседних элементов изображения. В этом случае кодовое слово последующего пикселя изменяется в зависимости от оптической плотности предыдущего пикселя.

Методика построения тонового кода состоит в следующем. Определяются условные вероятности $P(i|j)$, где j – оптическая плотность предыдущего элемента. Затем для каждого значения j строится свой эффективный код для последующего элемента. Это означает, что при одной и той же оптической плотности последующего элемента его кодовая комбинация будет различной и зависимой от оптической плотности предыдущего элемента.

Недостатки такого метода кодирования – необходимость предварительного изучения характера изображений, т. е. адаптации системы кодирования, и отсутствие помехоустойчивости. Искажение одного отсчета приводит к искажению всех последующих.

11.3.2. Кодирование с образованием блоков

Кодирование с образованием блоков состоит в том, что изображение разрезается на блоки, как правило, квадратные, в пределах которых осуществляется раздельное кодирование математического ожидания оптической плотности и его случайной составляющей.

Математическое ожидание и дисперсия определяются формулами

$$m_x = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x(i,j),$$

$$D_x = \frac{1}{(n-1)^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x^2(i,j) - m_x^2.$$

где x – оптическая плотность.

Обычно блоки принимают размерами 4×4 пикселей. Для каждого пикселя информация в один бит обозначает, находится ли уровень оптической плотности пикселя выше или ниже среднего. Значения m и D вычисляются отдельно для каждого блока.

При реконструкции исходного изображения значения уровней квантования принимаются в соответствии со следующими формулами:

$$x_0 = m_x - \sqrt{D_x \frac{n_1}{n_2}};$$

$$x_1 = m_x - \sqrt{D_x \frac{n_1}{n_2}},$$

где n_1 и n_2 – число пикселей в блоке, оптическая плотность которых выше или ниже значения m_x. Пиксели, закодированные единицами, приравниваются к значению x_1, а закодированные нулями – к x_0.

Достоинство этого метода – его простота, а также значительное сжатие данных при кодировании изображений.

11.3.3. Кодирование с предсказанием отсчетов

Кодирование с предсказанием основано на том, что каждый отсчет оптической плотности x_i можно предсказать на основе анализа нескольких предыдущих отсчетов, x_{i-1}, x_{i-2}, x_{i-3},... Этот принцип базируется на взаимозависимости корреляции между оптическими плотностями соседних пикселей.

При правильно выбранной функции предсказания объем памяти для представления изображения сокращается в 3–3,5 раза.

Эта функция зависит не только от координат α_i (номера пикселя), но и от некоторых параметров a, b, c:

$$x = (\varphi(\alpha; a, b, c)).$$

Если учитывать значения отсчетов при n пикселях, то условием оптимальности выбора функции является минимум среднего квадрата отклонения оптической плотности от ее аппроксимирующей функциональной зависимости:

$$\sum_{i=1}^{n}[x_i - \varphi(\alpha_i; a, b, c)]^2 = \min .$$

Из этого условия можно найти или подобрать параметры a, b и c, для чего необходимо дифференцировать апроксимирующую функцию по этим параметрам.

Параметры функции предсказания ($\varphi(\alpha; a, b, c)$) можно представить в виде полинома с параболической зависимостью $x = a\alpha^2 + b\beta + c$ или, например, в виде $x = a\alpha + b\beta + c$, где экстраполяция оптической плотности пикселей может производиться по строке и по кадру.

11.4. Кодирование штриховых изображений

Первичное описание штриховых изображений также связано с использованием растрового формата представления информации, так как он органически связан с принципом функционирования сканеров при вводе изображений в ЭВМ. Однако в связи со специфичностью штриховых изображений методы их вторичного описания могут существенно отличаться от методов описания тоновых изображений.

В штриховом изображении (при обработке чертежей, схем, карт) число градаций оптической плотности равно двум, поэтому оптическая плотность кодируется только двумя символами: 0 или 1. Этот способ кодирования обладает большей компактностью по сравнению с представлением тоновых изображений. В принципе способы кодирования тоновых изображений пригодны и для двухградационных изображений. Но уменьшение числа градаций до двух позволяет использовать и другие, более эффективные методы кодирования.

11.4.1. Кодирование длинами серий

Кодирование длинами серий состоит в представлении каждой строки растра в виде последовательности числа пикселей, составляющих эту строку, нулевых или единичных серий. Для представления длин серий часто используется префиксный алгоритм Хаффмена, в котором ни одно из кодовых слов не может быть начальной частью

(префиксом) другого. Это позволяет определять границы кодовых слов без использования дополнительных граничных кодов.

Эффективное кодирование длинами серий возможно при знании плотности распределения этих длин. Если l_0 – длины нулевых серий, а l_1 – длины единичных серий, то энтропия их будет

$$H(l_0) = -\sum_{i=1}^{n} P_{0i} \log_2 P_{0i},$$

$$H(h) = -\sum_{i=1}^{n} P_{1i} \log_2 P_{1i}.$$

Вероятности появления определенных длин серий 0 и 1 различны, а поэтому и их энтропии тоже будут различными. В связи с этим кодирование длин серий 0 и 1 будет неодинаковым. При анализе чертежей и схем длины линий, несущих информацию о темных участках изображения, будут значительно меньше, чем длины линий светлых участков.

Эффективное кодирование, при котором длина кодового слова равна его энтропии, обеспечивает наибольшее сжатие данных. В этом случае объем данных об изображении сжимается в 5–7 раз. К сожалению, неравномерность такого кода существенно усложняет процесс обработки изображения. Поэтому оно кодируется словами одинаковой длины.

Если кодовое слово содержит m разрядов, то с его помощью можно закодировать N отсчетов. При этом должно выполняться условие $n \le 2^m$.

Для определения коэффициента сжатия информации представим, что появление нулей и единиц в строке представляет собой Марковский процесс с переходными вероятностями взаимосвязи событий $P(i|j)$:

$$\mathrm{P} = \begin{matrix} 0 \\ 1 \end{matrix} \left\| \begin{matrix} 0 & 1 \\ \mathrm{P}_{00} & 1-\mathrm{P}_{00} \\ 1-\mathrm{P}_{11} & \mathrm{P}_{11} \end{matrix} \right\| .$$

Средние длины серий будут равны

$$m_{l_0} = \frac{1}{1-P_{00}},\ m_{l_1} = \frac{1}{1-P_{11}}.$$

При длине кодового слова, равном m, коэффициент сжатия данных равен

$$k = \frac{m_{l_0} + m_{l_1}}{2m}.$$

Расчеты показывают, что этот коэффициент обычно находится в пределах от 3 до 5.

11.4.2. Кодирование блоками

Кодирование блоками заключается в том, что изображение разбивается на отдельные прямоугольные блоки, которые кодируются в соответствии с вероятностями появления на них светлых пикселей. Наибольшего сжатия данных в этом случае можно достигнуть при использовании алгоритма Хаффмена, но его неравномерность затрудняет декодирование информации.

Анализ штриховых изображений показывает, что наиболее вероятными в них являются светлые пиксели. Для кодирования таких изображений целесообразно применять метод кодирования с пропуском белого. Если вероятность того, что блок размером $l_r \times l_b$ пикселей состоит только из нулевых элементов, $P(0;\ l_r,\ l_b)$, то средняя длина кодового слова будет

$$m_m = P(0;\ l_r,\ l_b) + (1 + l_r l_b)[1 - P(0;\ l_r,\ l_b)].$$

Откуда

$$m_m = l_r l_b [1 - P(0;\ l_r,\ l_b)] + 1.$$

Коэффициент сжатия данных при этом равен

$$k = \frac{1}{P(0;\ l_r,\ l_b) + \dfrac{1}{l_r l_b}}.$$

В случае, когда $l_b = 1$, блок принадлежит одной строке и код называется одномерным, в противном случае он называется двухмерным. Если ввести ранее принятые в матрице обозначения $P(1|0) = 1 - P_{11}$, $P(0|1) = 1 - P_{00}$, то вероятность наличия нулей в строке будет

$$P(0;\ l_r,\ 1) = \frac{1 - P_{00}}{2 - P_{00} - P_{11}} P_{11}^{l_{2-1}}.$$

Коэффициент сжатия данных для одномерных кодов изменяется в пределах от 2 до 5, а для двухмерных – от 4 до 8.

11.4.3. Векторное кодирование

Векторное кодирование применяется в основном для описания чертежей, в которых содержащаяся в них информация может быть представлена в виде набора линий. Метод векторного кодирования предполагает, что каждая линия может быть аппроксимирована ломаной, которая задается граничными точками и шириной.

Векторные форматы предусматривают использование данных о координатах угловых точек, толщине и цвете контурных линий, типе и цвете сплошной заливки участков изображения. Фиксируется также относительное расположение объектов и элементов изображения на странице.

Метод выделения векторов рассчитан на его применение при сканировании изображения. При этом принадлежность серии пикселей от темных участков чертежа к одному отрезку прямой линии определяется сравнением соседних строк растра. Когда образуется вектор, координаты левой границы серии на начальной линии сканирования принимаются за координаты начала вектора, а в качестве координат конца вектора принимаются координаты левой границы последней серии. Ширина отрезка прямой линии, представляемая в виде вектора, принимается равной длине серии пикселей.

Уравнение отрезка прямой линии представляется в обычной форме:

$$\frac{\beta - \beta_1}{\beta_2 - \beta_1} = \frac{\alpha - \alpha_1}{\alpha_2 - \alpha_1},$$

где α_1, β_1; α_2, β_2 – координаты начала и конца отрезка прямой линии, или же в форме

$$\beta = k\alpha + b,$$

где $k = \mathrm{tg}\,\phi$ – угловой коэффициент прямой.

Векторное кодирование может дать сжатие данных в 40 раз. При таком способе кодирования происходит потеря информации, которая обычно несущественна при обработке чертежей. За счет этого и обеспечивается сжатие данных об изображении.

11.4.4. Кодирование методом аппроксимации

Кодирование методом аппроксимации известными функциями применяется тогда, когда характер линий не позволяет представить их в виде совокупности отрезков прямых.

При выборе метода аппроксимации к аппроксимирующим функциям может быть предъявлено одно из следующих требований:

– полное совпадение исходной и аппроксимирующей функций в узловых точках;

– минимальное значение суммы квадратов отклонений в узловых точках;

– минимальное значение суммы квадратов максимальных отклонений аппроксимирующей функции от исходной.

Недостатком использования полиномов для аппроксимации линий является некоторая волнистость у полиноминальных функций высокого порядка. Поэтому для описания некоторых линий, где такая волнистость нежелательна, применяются сплайн-функции. Они обладают весьма высокой степенью гладкости и используются для аппроксимации контуров знаков при проектировании шрифтов – это так называемые кривые Безье.

Лекция 13–14. КОДИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

Текстовая информация в исходном состоянии состоит из конечного числа символов (букв, цифр, знаков препинания, пробелов, математических знаков). Эта совокупность знаков образует первичный алфавит.

Первичный алфавит состоит из большого числа знаков, поэтому его использование при образовании информации крайне затруднительно. Для удобства переработки информации необходимо преобразовать ее с помощью вторичного алфавита, число знаков которого невелико. Алгоритм преобразования первичного алфавита во вторичный называется кодом, а операция такого преобразования – кодированием. Операция обратного преобразования вторичного алфавита в первичный называется декодированием. Последовательность сигналов вторичного алфавита, которая соответствует содержанию передаваемого сообщения, называют кодовым словом.

В процессе кодирования информации при отсутствии помех можно достичь следующих целей:

– обеспечить простоту, надежность и эффективность аппаратных устройств информационных систем;

– минимизировать время передачи информации;

– минимизировать объем запоминающего устройства при хранении информации;

– обеспечить простоту выполнения арифметических и логических операций.

При передаче информации в условиях помех важнейшее значение приобретает достоверность распознавания сообщений. Заданная достоверность обычно обеспечивается внесением избыточности сообщения, позволяющей исключить ошибки передачи сигналов.

13.1. Оптимальные системы счисления

Найдем оптимальную систему счисления с использованием критерия минимума элементов, которые служат для определения максимального числа. Пусть k – основание системы счисления, тогда для изображения n-разрядного числа необходимо использовать $v = kn$ элементов. Максимальное число, которое можно представить при системе счисления с основанием k и n разрядах, равно

$$N_{\max} = k^n - 1. \tag{13.1}$$

Найдем n из выражения (13.1):

$$n \ln k = \ln(N_{\max} + 1),$$

$$n = \frac{\ln(N_{\max} + 1)}{\ln k},$$

следовательно,

$$v = \ln(N_{\max} + 1)\frac{k}{\ln k}. \tag{13.2}$$

Определим, при каком значении k получается наименьшее значение v. Для этого продифференцируем выражение (13.2):

$$\frac{dv}{dk} = \ln(N_{\max} + 1)\frac{\ln k - 1}{(\ln k)^2}. \tag{13.3}$$

Приравняем к нулю выражение (13.3) и найдем:

$$\ln k - 1 = 0,$$

где $k = e = 2{,}718$ – основание натурального логарифма.

Следовательно, оптимальной системой счисления будет основание натурального логарифма.

Однако система счисления с дробным основанием с точки зрения технической реализации неприемлема. Поэтому выбирают основание $k = 2$, достаточно близкое к основанию оптимальной системы счисления. При этом увеличение числа элементов для представлений максимального числа будет незначительным (6%).

Цифры двоичного кода позволяют легко осуществлять арифметические и логические операции.

Однако в связи с трудностью и непривычностью оценки оператором двоичного числа получили распространение и другие системы счисления, которые сводятся как к двоичной, так и к десятичной системе счисления. Так существуют восьмеричная и двоично-десятичная системы счисления. В восьмеричной системе для записи всех возможных чисел используется восемь чисел от 0 до 7. Перевод из восьмеричной системы в двоичную осуществляется заменой восьмеричной цифры равным трехразрядным числом.

Например, 714 запишем 111 001 100 (111 = 7, 001 = 1, 100 = 4).

Двоично-десятичная система счисления сохраняет преимущества и двоичной, и десятичной систем счисления. При этом каждую цифру

десятичного числа записывают в виде четырехразрядного двоичного числа. Этот код используется в качестве промежуточного при вводе в ЭВМ данных, представленных в десятичном коде. Один из вариантов такого кода представлен в табл. 13.1.

Таблица 13.1

Десятичный код	Двоично-десятичный код	Десятичный код	Двоично-десятичный код
0	0000 0000	6	0000 0110
1	0000 0001	7	0000 0111
2	0000 0010	8	0000 1000
3	0000 0011	9	0000 1001
4	0000 0100	10	0001 0000
5	0000 0101		

Однако двоичный код в силу ряда причин, а именно сжатия информации, простоты выполнения в этой системе арифметических и логических операций, занял особое место и широко используется при кодировании текстовой информации.

13.2. Основные параметры кодов

При кодировании текстовой информации возникают довольно сложные проблемы. Необходимо учитывать как вероятности появления букв, так и возможности обнаружения ошибок передачи.

Кодовые комбинации, используемые для представления заданного количества сообщений, называются кодовым **словарем.** Важным показателем кода является его основание, которое равно основанию выбранной системы счисления. Основание определяет число различных символов, с помощью которых образуются кодовые комбинации. Эти символы называют **элементами кода.** Для передачи сообщения символы вторичного алфавита-кода необходимо разместить на определенных местах-разрядах. Число разрядов n в коде образуют длину кода.

Если длина кода – n разрядов, то можно этим кодом передать $A = 2^n$ сообщений. В двоичной системе каждый символ первичного алфавита является сообщением. Поэтому для передачи символов число разрядов кода (длина кода) определяется по формуле

$$N = \log_2 A .$$

Например, для передачи 128 сообщений нужен код длиной $n = \log_2 128 = 7$. При кодировании используются равномерные и не-

равномерные коды. У равномерных кодов длина всех сообщений одинакова, а у неравномерных, для разных сообщений – разная. Примером равномерного кода является код Бодо (пятиэлементный), а неравномерного – код Морзе.

Различают последовательные и параллельные коды. Это понятие относится к способам ввода, вывода кодированной информации и передачи ее по каналам связи. В том случае, когда все разряды кода вводятся, выводятся и передаются одновременно, коды называют **параллельными**. Если же ввод, передача и вывод кодовой комбинации осуществляются последовательно разряд за разрядом, то такие коды называются **последовательными.**

Параллельные коды требуют меньшего времени на передачу, но при этом увеличивается число линий связи. При последовательных кодах линия связи одна, но время передачи увеличивается.

13.3. Способы представления кодов

Любое число или код, независимо от системы счисления, может быть представлено многочленом следующего вида:

$$N = a_{\text{И}-1}k^{n-1} + ... + a_3k^3 + a_2k^2 + a_1k^1 + a_0k^0,$$

где n – число разрядов, k – основание системы счисления, a_{n-1}, …, a_3, a_2, a_1, a_0 – коэффициенты, принимающие значение от 0 до $k-1$. В десятичной системе счисления $k = 10$, а a_0, a_1, a_2, … – цифры от 0 до 9. Поэтому, например, число 1245 запишется следующим образом:

$$1245 = 1 \cdot 10^3 + 2 \cdot 10^2 + 4 \cdot 10^1 + 5 \cdot 10^0.$$

В двоичной системе счисления $k = 2$, a_0 и a_1 – коэффициенты. Число 10 111 в этой системе счисления может быть представлено так:

$$10111 = 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0.$$

Одно и то же число может быть записано в различных системах счисления. Например, десятичное число 125 в двоичной системе счисления запишется следующим образом:

$$125 = 1 \cdot 2^8 + 1 \cdot 2^7 + 1 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^3 + 1 \cdot 2^1 + 1 \cdot 2^0 = 1111101.$$

При кодировании символьной информации основными требованиями являются общее число кодируемых сообщений и возможность обнаруживать искажения информации. Для этого разработаны специальные коды, некоторые из них будут рассмотрены далее.

13.4. Эффективное кодирование

Эффективное кодирование состоит в том, чтобы, учитывая статистические свойства источника сообщения (вероятность появления каждого знака первичного алфавита), минимизировать среднее число двоичных знаков, используемых для кодирования одного знака. В этом случае время на передачу сообщения, а также объем запоминающего устройства минимальны.

Теоретическая база эффективного кодирования – теорема, доказанная К. Шенноном для каналов передач в отсутствие помех (шумов).

Теорема Шеннона. Сообщения, составленные из букв некоторого алфавита, можно закодировать так, что среднее число двоичных символов на букву будет сколь угодно близким к энтропии источника сообщений, но не меньше этой величины.

Теорема не указывает конкретный способ кодирования, но из нее следует, что при выборе каждого символа необходимо стремиться к тому, чтобы он нес максимальную информацию.

Конструктивные методы построения для случая отсутствия статистической взаимосвязи между буквами были даны независимо К. Шенноном и Р. Фано. Их методы отличаются незначительно, поэтому соответствующий код получил название кода Шеннона – Фано.

Метод построения кода заключается в следующем:

– буквы алфавита выписываются в таблицу в порядке убывания вероятностей $p(z_i)$;

– затем все буквы разделяются на две группы так, чтобы суммы вероятностей в каждой группе были примерно одинаковыми;

– всем буквам первой группы в качестве первого символа присваивается 1, а всем буквам второй группы – 0;

– каждая из полученных групп, в свою очередь, разбивается на две подгруппы так, чтобы суммы вероятностей у них были примерно одинаковыми и т. д.;

– всем буквам первой группы присваивается 1, а второй – 0;

– процесс продолжается до тех пор, пока в каждой подгруппе не останется по одной букве.

При эффективном кодировании наибольший эффект получается в том случае, когда вероятности определяются формулой

$$p_i = \left(\frac{1}{2}\right)^i,$$

где $i = 1, 2, 3\ldots$

Среднее число символов на букву сообщения в этом случае точно равно энтропии.

Рассмотрим пример построения оптимального кода по методике Шеннона-Фано при произвольных значениях вероятности появления букв.

Пример. Исходный алфавит состоит из восьми букв, вероятности появления которых показаны в табл. 13.2.

Таблица 13.2

Буква	P	Буква	P	Буква	P	Буква	P
Z1	0,22	Z3	0,16	Z5	0,10	Z7	0,04
Z2	0,20	Z4	0,16	Z6	0,10	Z8	0,02

Результаты кодирования представлены в табл. 13.3.

Таблица 13.3

Буква	Вероятность	Кодовые комбинации	Ступени разбиения
*Z*1	0,22	11	I
*Z*2	0,20	101	II
*Z*3	0,16	100	III
*Z*4	0,16	01	IV
*Z*5	0,10	001	V
*Z*6	0,10	0001	VI
*Z*7	0,04	00001	VII
*Z*8	0,02	00000	VIII

В данном примере первое разбиение делит все буквы на две группы: *Z*1–*Z*3, *Z*4–*Z*8. Второе разбиение выполняется для первой группы. Им она делится на две подгруппы: *Z*1и *Z*2–*Z*3; третье разделение делит вторую подгруппу еще раз: *Z*2 и *Z*3; четвертое разбиение делит вторую группу на две подгруппы: *Z*4 и *Z*5–*Z*8; пятое разбиение – на *Z*5 и *Z*6–*Z*8; шестое – на *Z*6 и *Z*7–*Z*8; седьмое – на *Z*7 и *Z*8.

Вычислим:

$$H(z_i) = -\sum_{i=1}^{8} p(z_i) \log_2 p(z_i),$$

и среднее число символов:

$$l_{cp} = \sum_{i=1}^{m} p(z_i) n(z_i),$$

где m – число букв, n – число двоичных символов в коде i-й буквы, дают следующие результаты:

$$H(z_i) = 2{,}76\,,\ l_{\text{ср}} = 2{,}84\,.$$

Таким образом, при произвольном распределении вероятностей букв среднее число символов на букву становится больше энтропии, но меньше трех, определяемых по формуле.

Среднее число символов на букву $l_{\text{ср}} = 2{,}84$, что близко к значению энтропии.

Рассмотренный пример показывает, что при использовании кода Шеннона – Фано избыточность остается. Ее можно уменьшить, если перейти от кодирования букв к кодированию блоков, составленных из букв алфавита. Повышение эффективности кодирования при переходе к блокам объясняется тем, что в этом случае удается получить группы с более близкими вероятностями.

13.5. Код Хаффмена

Методика Хаффмена обеспечивает однозначное построение кода с наименьшим для данного распределения вероятностей средним числом символов на букву. Для кода с основанием 2 она сводится к следующему.

Буквы алфавита выписываются в основной столбец таблицы в порядке убывания вероятностей. Две последние буквы объединяются в одну вспомогательную, которой приписывается суммарная вероятность. Вероятности букв, не участвовавших в объединении, и полученная суммарная вероятность вновь располагаются в порядке убывания во вспомогательном столбце. Две последние вероятности вновь объединяются, и заполняется второй вспомогательный столбец и т. д.

Процесс продолжается до тех пор, пока не получится вспомогательная буква с вероятностью, равной 1. На этом заканчивается первый этап. Поясним методику построения на примере.

Пример. Задан алфавит из восьми букв: $Z1$, $Z2$, $Z3$, $Z4$, $Z5$, $Z6$, $Z7$, $Z8$ с вероятностями:

$$p_1 = 0{,}22,\ p_2 = 0{,}20,\ p_3 = 0{,}16,\ p_4 = 0{,}16,$$
$$p_5 = 0{,}10,\ p_6 = 0{,}10,\ p_7 = 0{,}04,\ p_8 = 0{,}02.$$

Методика выполнения первого этапа иллюстрируется в табл. 13.4.

Таблица 13.4

Буква	Вероятность	Вспомогательные столбцы						
		1	2	3	4	5	6	7
*Z*1	0,22	0,22	0,22	0,26	0,32	0,42	0,58	1
*Z*2	0,20	0,20	0,20	0,22	0,26	0,32	0,42	
*Z*3	0,16	0,16	0,16	0,20	0,22	0,26		
*Z*4	0,16	0,16	0,16	0,16	0,20			
*Z*5	0,10	0,10	0,16	0,16				
*Z*6	0,10	0,10	0,10					
*Z*7	0,04	0,06						
*Z*8	0,02							

На втором этапе строится кодовое дерево. Для составления кодовой комбинации необходимо проследить путь перехода по строкам и столбцам таблицы. Эта задача облегчается при использовании кодового дерева. Из точки, соответствующей вероятности 1, направляются две ветви. Ветви с большей вероятностью присваивается символ 1, а с меньшей – символ 0. Такое последовательное ветвление продолжается до тех пор, пока ветвь не дойдет до каждой буквы.

На третьем этапе по кодовому дереву составляется кодовая комбинация. Двигаясь по кодовому дереву от 1, можно записать для каждой буквы соответствующую ей кодовую комбинацию (см. табл. 13.5). Определяя $l_{ср}$, получим:

$$l_{ср} = 2 \cdot 0{,}22 + 2 \cdot 0{,}20 + 3 \cdot 0{,}16 + 3 \cdot 0{,}16 + 3 \cdot 0{,}10 + 4 \cdot 0{,}10 + 5 \cdot 0{,}04 + \\ + 5 \cdot 0{,}02 = 0{,}44 + 0{,}40 + 0{,}48 + 0{,}48 + 0{,}30 + 0{,}40 + 0{,}20 + 0{,}10 = 2{,}8.$$

Таблица 13.5

Буква	Код	Буква	Код	Буква	Код	Буква	Код
*Z*1	01	*Z*3	111	*Z*5	100	*Z*7	10101
*Z*2	00	*Z*4	110	*Z*6	1011	*Z*8	10100

Все эффективные коды являются неравномерными, так как буквам с большими значениями вероятностей их появления присваиваются кодовые комбинации меньшей длины.

Лекция 15. ПОМЕХОУСТОЙЧИВОЕ КОДИРОВАНИЕ

15.1. Пути совершенствования эффективного кодирования

При эффективном кодировании среднее количество знаков приближается к значению энтропии за счет присвоения более вероятным знакам первичного алфавита более коротких кодовых комбинаций и более длинных комбинаций – менее вероятным знакам. Неравномерность кодов существенно усложняет задачу декодирования, так как при этом трудно найти начало и конец символа. Введение разделительного символа весьма невыгодно, так как средняя длина кодовой комбинации при этом возрастает.

Более рациональным методом разделения знаков является такое построение кода, при котором ни одна комбинация кода не совпадает с началом более длинной комбинации. Такие коды называются **префиксными.**

Пусть задана последовательность кодов 1000001 101 10 10100, которые имеют вид: $x1 = 00$, $x2 = 01$, $x3 = 101$, $x4 = 100$. Эта последовательность декодируется однозначно: $x4(100)$, $x1(00)$, $x2(01)$, $x3(101)$, $x3(101)$, $x3(101)$, $x1(00)$. Таким образом, это последовательность префиксного кода.

Последовательность 000101010101 комбинаций непрефиксного кода $x1 = 00$, $x2 = 01$, $x3 = 101$, $x4 = 010$ может быть декодирована различными способами, так как комбинация 01 является началом комбинации 010.

Коды, полученные по методике Шеннона-Фано и Хаффмена, удовлетворяют требованиям префиксных кодов.

Для эффективного кодирования коррелированных последовательных знаков целесообразно укрупнить знаки первичного алфавита. При этом подлежащее передаче сообщение разбивается на двух-, трех- или n-знаковые сочетания, вероятности которых определяются на основе вероятностей появления знаков исходного алфавита. Каждому такому сочетанию ставится в соответствие кодовая комбинация, найденная по одной из изложенных методик.

При таком объединении знаков первичного алфавита кодирование облегчается, т. к. корреляция между укрупненными знаками существенно ослабляется.

Если в канале связи отсутствуют помехи, то при приеме сигнала, зная его код, можно точно установить, что это за сигнал и какую информацию он содержит. При наличии помех возможно искажение информации. Следовательно, необходимо принимать специальные меры для того, чтобы повысить достоверность принятой информации. Главными среди этих мер являются следующие:

– увеличение мощности передаваемого сигнала с тем, чтобы мощность полезного сигнала значительно превышала мощность помехи;

– повторение кодовых комбинаций;

– помехоустойчивое кодирование.

Увеличение мощности передаваемого сигнала связано с энергетическими затратами, а повторение передаваемой информации увеличивает время и дополнительные затраты энергии на ее передачу. Поэтому основным средством борьбы с влиянием помех на достоверность приема информации следует считать помехоустойчивое кодирование.

Теория помехоустойчивого кодирования базируется на результатах исследований К. Шеннона.

Теорема Шеннона. При любой скорости передачи двоичных сигналов, меньшей, чем пропускная способность канала связи, существует такой код, при котором вероятность ошибочного декодирования может быть сделана произвольно малой; если скорость передачи больше пропускной способности канала связи, то вероятность ошибки не может быть сделана произвольно малой.

На основе теории помехоустойчивого кодирования разработаны помехоустойчивые коды (их также называют корректирующими). Эти коды позволяют не только обнаруживать ошибки, но и исправлять их.

Рассмотрим двоичный код длиной n. С помощью этого кода можно получить $N = 2^n$ комбинаций. Ошибка при приеме состоит в том, что в результате действия помехи вместо 1 принят 0 или вместо 0 принята 1. Если в кодовой комбинации один знак заменяется другим, то ошибка называется одиночной, если два – двойной, если три – тройной и т. д.

Если при передаче используются все N комбинаций, то ошибка любой кратности остается незамеченной, так как при этом одна из возможных комбинаций переходит в другую. Но любая кодовая комбинация является допустимой – разрешенной. Поэтому с полным основанием принятую комбинацию можно считать верной, хотя на самом деле она является неверной.

Рассмотрим пример. Пусть четыре разных сообщения $Z1$, $Z2$, $Z3$ и $Z4$ закодированы двухзначным кодом: $n = 2$; $N = 2^n = 2^2 = 4$.

Следовательно, передаваемые сообщения могут быть закодированы так:

Буква	$Z1$	$Z2$	$Z3$	$Z4$
Код	00	01	10	11

Если передается сообщение $Z2$, то ошибка при приеме в первом разряде переводит комбинацию 01 в комбинацию 11, то есть вместо сообщения $Z2$ принимается сообщение $Z4$. Ошибка во втором разряде переводит комбинацию 01 в комбинацию 00, и вместо сообщения $Z2$ будет принято сообщение $Z1$ и т. д.

Это происходит потому, что комбинации 00 и 01, 10 и 11, 01 и 11 различаются только в одном знаке. Для того чтобы можно было обнаружить одиночную ошибку необходимо, чтобы комбинации между собой различались не менее, чем в двух знаках. В этом случае одиночная ошибка даст комбинацию, которая от истинной и от любой другой будет отличаться в одном или двух знаках. На этом основан принцип построения кода, обнаруживающего ошибку. Принцип формулируется так: правило построения кода, позволяющего обнаруживать одиночную ошибку, заключается в том, что из всех возможных комбинаций используется только половина.

Пример. Необходимо для передачи сообщений $Z1$, $Z2$, $Z3$, $Z4$ построить код, позволяющий обнаруживать одиночную ошибку.

Для построения кода нужно, чтобы выбранные четыре комбинации отличались друг от друга не менее, чем в двух разрядах. Это возможно, если для кодирования использовать трехразрядные кодовые комбинации, но только половину из них. Всего комбинаций при $n = 3$ равно $N = 2^3 = 8$.

Поэтому для кодирования выберем только четыре:

Буква	$Z1$	$Z2$	$Z3$	$Z4$
Код	000	011	101	110

Остальные четыре комбинации 001, 010, 100, 111 будут запрещенными. Пусть при приеме сообщения $Z2$ произошла ошибка во втором разряде, и вместо комбинации 011 принята комбинация 001. Эта комбинация относится к запрещенным и, следовательно, является ошибочной.

В рассмотренном примере устанавливается только факт ошибки, но нет ответа на вопрос, какая из четырех комбинаций была передана.

Для того чтобы не только установить наличие ошибки, но и указать, какая комбинация передавалась, необходимо построить код, позволяющий исправить ошибку. В таком коде все кодовые комбинации должны отличаться не менее, чем в трех разрядах. Для этого кода одиночная ошибка дает кодовую комбинацию, отличающуюся от истинной в одном разряде, а от ближайшей разрешенной – не менее, чем в двух разрядах.

Пример. Закодировать четыре сообщения *Z*1*, Z*2*, Z*3 и *Z*4 так, чтобы можно было обнаруживать и исправлять одиночную ошибку.

Для кодирования необходимо выбрать пятиразрядный код и из всех комбинаций отобрать только отличающиеся друг от друга не менее, чем в трех разрядах:

Буква	*Z*1	*Z*2	*Z*3	*Z*4
Код	00000	01101	10110	11011

Пусть вместо переданной комбинации *Z*2 = 01101 принята комбинация 01001. Такой комбинации среди разрешенных нет. Поэтому сравним эту комбинацию со всеми разрешенными. В результате сравнения получим: от *Z*1 отличие в двух знаках, от Z3 отличие в пяти знаках, от *Z*4 отличие в двух знаках, а от *Z*2 отличие в одном знаке. Следовательно, была передана комбинация *Z*2, и необходимо принятую комбинацию исправить, то есть изменить во втором разряде 0 на 1.

15.2. Основные характеристики корректирующих кодов

В настоящее время наибольшее внимание с точки зрения технических приложений уделяется двоичным блочным корректирующим кодам.

При общем числе n символов в блоке число информационных символов равно k, а число проверочных символов

$$r = n - k .$$

К основным характеристикам корректирующих кодов относятся:

– число разрешенных и запрещенных кодовых комбинаций;

– избыточность кода;

– минимальное кодовое расстояние;

– число обнаруживаемых и исправляемых ошибок;

– корректирующая возможность кодов.

Для блочных двоичных кодов с числом символов в блоках, равным n, общее число возможных кодовых комбинаций определяется значением

$$N = 2^n.$$

Число разрешенных кодовых комбинаций при наличии k информационных разрядов в первичном коде равно

$$N_k = 2^k.$$

Очевидно, что число запрещенных комбинаций равно

$$N_з = N_0 - N_k = 2^n - 2^k,$$

отсюда можно записать:

$$N_0 / N_k = 2^n / 2^k = 2^{n-k} = 2^r .$$

где r – число избыточных (проверочных) разрядов в блочном коде.

Избыточностью корректирующего кода называют величину

$$X = \frac{r}{n} = \frac{n-k}{n} = 1 - \frac{k}{n}.$$

Скорость передачи информации после кодирования ее корректирующим кодом определяется формулой

$$B = H \cdot k / n,$$

где H – производительность источника информации, символов в секунду.

Если число ошибок, которые нужно обнаружить или исправить, значительно, то необходимо иметь код с большим числом проверочных символов. Чтобы при этом скорость передачи оставалась достаточно высокой, необходимо в каждом кодовом блоке одновременно увеличивать как общее число символов, так и число информационных символов. При этом длительность кодовых блоков будет существенно возрастать, что приведет к задержке информации при передаче и приеме. Чем сложнее кодирование, тем длительнее временная задержка информации.

Для того чтобы можно было обнаружить и исправить ошибки, разрешенная комбинация должна как можно больше отличаться от запрещенной. Если ошибки в канале связи действуют независимо, то вероятность преобразования одной кодовой комбинации в другую будет тем меньше, чем большим числом символов они различаются.

Если интерпретировать кодовые комбинации как точки в пространстве, то отличие выражается в близости этих точек, т. е. в расстоянии между ними.

Количество разрядов (символов), которыми отличаются две кодовые комбинации, можно принять за кодовое расстояние между ними. Для определения этого расстояния нужно сложить две кодовые комбинации по модулю 2 и подсчитать число единиц в полученной сум-

ме. Например, две кодовые комбинации $X_i = 01011$ и $X_j = 10010$ имеют расстояние $d(X_iX_j)$, равное 3, так как:

$$X_i = 01011 \rightarrow W = 3,$$

$$X_j = 10010 \rightarrow W = 2,$$

$$X_i + X_j = 11001 \rightarrow d\left(X_iX_j\right)W = 3.$$

Здесь под операцией «+» понимается сложение по mod2.

Заметим, что кодовое расстояние $d(X_iX_j)$ между комбинацией X_i и нулевой $X_0 = 00\ldots0$ называют весом W комбинации X_i, т. е. вес X равен числу «1» в ней.

Расстояние между различными комбинациями некоторого конкретного кода могут существенно отличаться. Так, в частности, в безызбыточном первичном натуральном коде ($n = k$) это расстояние для различных комбинаций может изменяться от единицы до величины n, равной значности кода. Особую важность для характеристики корректирующих свойств кода имеет минимальное кодовое расстояние $d_{\min}$, определяемое при попарном сравнении всех кодовых комбинаций и называемое расстоянием Хемминга.

В безызбыточном коде все комбинации являются разрешенными, и, следовательно, его минимальное кодовое расстояние равно единице $–d_{min} = 1$. Поэтому достаточно исказиться одному символу, чтобы вместо переданной комбинации была принята разрешенная комбинация.

Чтобы код обладал корректирующими свойствами, необходимо ввести в него некоторую избыточность, которая обеспечивала бы минимальное расстояние между любыми двумя разрешенными комбинациями не менее двух $–d_{\min} \geq 2$.

Минимальное кодовое расстояние – важнейшая характеристика помехоустойчивых кодов, указывающая на число обнаруживаемых или исправляемых заданным кодом ошибок.

При применении двоичных кодов учитывают только дискретные искажения, при которых единица переходит в нуль (1→0) или нуль переходит в единицу (0→1). Переход 1→0 или 0→1 только в одном элементе кодовой комбинации называют единичной ошибкой (единичным искажением). В общем случае под кратностью ошибки подразумевают число позиций кодовой комбинации, на которых под действием помехи одни символы оказались замененными на другие. Возможны двукратные ($g = 2$) и многократные ($g > 2$) искажения элементов в кодовой комбинации в пределах $0 < g < n$.

Минимальное кодовое расстояние – основной параметр, характеризующий корректирующие способности данного кода. Если код используется только для обнаружения ошибок кратностью g_1, то необходимо и достаточно, чтобы минимальное кодовое расстояние было равно

$$d_{\min} \geq g_0 + 1 .$$

В этом случае никакая комбинация из g_1 ошибок не может перевести одну разрешенную кодовую комбинацию в другую разрешенную. Таким образом, условие обнаружения всех ошибок кратностью g_0 можно записать в виде

$$g_0 \leq d_{\min} - 1 .$$

Чтобы можно было исправить все ошибки кратностью $g_и$ и менее, необходимо иметь минимальное расстояние, удовлетворяющее условию:

$$d_{\min} \geq g_и + 1 .$$

В этом случае любая кодовая комбинация с числом ошибок $g_и$ отличается от каждой разрешенной комбинации не менее, чем в $g_и + 1$ позициях. Если это условие не выполнено, возможен случай, когда ошибки кратности g исказят переданную комбинацию так, что она станет ближе к одной из разрешенных комбинаций, чем к переданной, или даже перейдет в другую разрешенную комбинацию. В соответствии с этим, условие исправления всех ошибок кратностью не более $g_и$ можно записать в виде

$$g_и \leq (d_{\min} - 1)/2 .$$

Из вышеприведенных формул следует, что если код исправляет все ошибки кратностью $g_и$, то число ошибок, которые он может обнаружить, равно $g_0 = 2g_и$.

Вопрос о минимально необходимой избыточности, при которой код обладает нужными корректирующими свойствами, является одним из важнейших в теории кодирования. В настоящее время получен лишь ряд верхних и нижних оценок (границ), которые устанавливают связь между максимально возможным минимальным расстоянием корректирующего кода и его избыточностью.

Так, для некоторых частных случаев Хемминг получил простые соотношения, позволяющие определить необходимое число проверочных символов:

$$r \geq \log_2(n+1) \text{ для } d_{\min} = 3 ,$$

$$r \geq \log_2(2n) \text{ для } d_{\min} = 4 .$$

Блочные коды с $d_{\min} = 3$ и $d_{\min} = 4$ в литературе называют кодами Хемминга.

Существующие методы построения избыточных кодов решают в основном задачу нахождения такого алгоритма кодирования и декодирования, который позволял бы наиболее просто построить и реализовать код с заданным значением $d_{\min}$. Поэтому различные корректирующие коды при одинаковых $d_{\min}$ сравниваются по сложности кодирующего и декодирующего устройств. Этот критерий является в ряде случаев определяющим при выборе того или иного кода.

15.3. Корректирующие коды Хемминга

Построение кодов Хемминга базируется на принципе проверки на четность веса W (числа единичных символов) в информационной группе кодового блока.

Критерием правильности таких комбинаций является равенство нулю результата S суммирования по mod2 всех n символов кода, включая проверочный символ. При наличии одиночной ошибки S принимает значение 1:

$$S = r_1 \oplus i_1 \oplus i_2 \oplus \ldots \oplus i_k = \begin{cases} 0 - \text{ошибки нет;} \\ 1 - \text{однократная ошибка.} \end{cases}$$

Этот код является $(k + 1, k)$ – кодом или $(n, n - 1)$ – кодом. Минимальное расстояние кода равно двум ($d_{\min} = 2$), и, следовательно, никакие ошибки не могут быть исправлены. Простой код с проверкой на четность может использоваться только для обнаружения (но не исправления) однократных ошибок.

Увеличивая число дополнительных проверочных разрядов и формируя по определенным правилам проверочные символы r, равные 0 или 1, можно усилить корректирующие свойства кода так, чтобы он позволял не только обнаруживать, но и исправлять ошибки. На этом и основано построение кодов Хемминга.

Рассмотрим эти коды, позволяющие исправлять одиночную ошибку с помощью непосредственного описания. Для каждого числа проверочных символов $r = 3, 4, 5\ldots$ существует классический код Хемминга с маркировкой

$$(n, k) = (2^r - 1, 2^r - 1 - r),$$

т. е. – (7, 4), (15, 11), (31, 26) … .

При других значениях числа информационных символов k получаются так называемые усеченные (укороченные) коды Хемминга. Так для международного телеграфного кода МТК-2, имеющего 5 информационных символов, потребуется использование корректирующего кода (9, 5), являющегося усеченным от классического кода Хемминга (15, 11), так как число символов в этом коде уменьшается (укорачивается) на 6. Для примера рассмотрим классический код Хемминга (7, 4). В простейшем варианте он может быть записан так: четыре ($k = 4$) заданных информационных символа (i_1, i_2, i_3, i_4), сгруппированные в начале кодового слова, дополняются тремя проверочными символами ($r = 3$). В результате получаем кодовые слова:

$$V = (i_1', \ i_2', \ i_3', \ i_4', \ r_1', \ r_2', \ r_3') \, .$$

Апостроф означает, что любой символ слова может быть искажен помехой в канале передачи.

Проверочные символы определяются следующими равенствами проверки:

$$r_1 = i_1 \oplus i_2 \oplus i_3 ;$$

$$r_2 = i_2 \oplus i_3 \oplus i_4 ;$$

$$r_3 = i_1 \oplus i_2 \oplus i_4 \, .$$

где знак $\oplus$ означает сложение по модулю 2.

В соответствии с этим алгоритмом определения значений проверочных символов ниже выписаны все возможные 16 кодовых слов (7, 4) – кода Хемминга:

$k = 4 \ | \ r = 3$

$i_1 \ i_2 \ i_3 \ i_4 \, | \, r_1 \ r_2 \ r_3$

———————

0000 000
0001 011
0010 110
0011 101
0100 111
0101 100
0110 001
0111 010
1000 101

1001 110
1010 011
1011 000
1100 010
1101 001
1110 100
1110 111

При декодировании в режиме исправления ошибок строится последовательность

$$s_1 = r_1' \oplus i_1' \oplus i_2' \oplus i_3';$$

$$s_2 = r_2' \oplus i_2' \oplus i_3' \oplus i_4';$$

$$s_3 = r_3' \oplus i_1' \oplus i_2' \oplus i_4'.$$

Трехсимвольная последовательность (S_1, S_2, S_3) называется синдромом.

В данном случае синдром $S = (S_1, S_2, S_3)$ представляет собой сочетание результатов проверки на четность соответствующих символов кодовой группы и характеризует определенную конфигурацию ошибок.

Число возможных синдромов определяется выражением

$$S = 2^r.$$

При числе проверочных символов $r = 3$ имеется восемь возможных синдромов ($2^3 = 8$). Нулевой синдром (000) указывает на то, что ошибки при приеме отсутствуют или не обнаружены. Всякому ненулевому синдрому соответствует определенная конфигурация ошибок, которая и исправляется. Классические коды Хемминга имеют число синдромов, точно равное их необходимому числу, позволяют исправить все однократные ошибки в любом информативном и проверочном символах и включают один нулевой синдром. Такие коды называются плотноупакованными.

Усеченные коды являются неплотно упакованными, так как число синдромов у них превышает необходимое. Так, в коде (9, 5) при четырех проверочных символах число синдромов будет равно $2^4 = 16$, в то время как необходимо всего 10. Лишние 6 синдромов свидетельствуют о неполной упаковке такого кода.

Для рассматриваемого кода (7, 4) в табл. 10 представлены ненулевые синдромы и соответствующие конфигурации ошибок.

Таблица 10

Синдром	001	010	011	100	101	110	111
Конфигу-рация ошибок	0000001	0000010	0001000	0000100	1000000	0010000	0100000
Ошибка в символе	r_3	r_2	i_4	r_1	i_1	i_3	i_2

Таким образом, код (7, 4) позволяет исправить все одиночные ошибки. Простая проверка показывает, что каждая из ошибок имеет свой единственный синдром. При этом возможно создание такого цифрового корректора ошибок (дешифратора синдрома), который по соответствующему синдрому исправляет соответствующий символ в кодовой группе. Две или более ошибки превышают возможности корректирующего кода Хемминга, и декодер будет ошибаться. Это означает, что он будет вносить неправильные исправления и выдавать искаженные информационные символы.

Лекция 16. ИНФОРМАЦИЯ И ЕЕ ВИДЫ

Информация является определяющим понятием кибернетики, а теория информации есть составная часть кибернетики как науки. Кибернетика – это наука о связях переработки информации и управления в технических системах, живой природе и обществе. Кибернетика как наука была провозглашена в 1948 г. американским математиком Винером.

В теории управления информация – количественная мера устранения неопределенности, мера организованности системы. С практической точки зрения информация (разъяснение, изложение) – это сведения, являющиеся объектом генерирования, хранения, передачи, преобразования, отображения и восприятия.

На этапах передачи информация проявляет себя в виде сигналов. Сигналом (от лат. *signum* – знак) называется процесс или явление, несущее сообщение о каком-либо событии, состоянии объекта или передающее команды управления, связи и оповещения. Следовательно, сигнал является материальным носителем информации. Понятие сигнала было впервые сформулировано в кибернетике.

По своей физической природе сигналы бывают механическими (перемещение и деформация объектов, изменение давления), тепловыми (изменение температуры), световыми (изменение силы света, освещенности, цвета изображения, зрительный образ), электромагнитными (радиоволны, рентгеновское излучение), звуковыми (акустические колебания), электрическими (изменение силы тока или напряжения – амплитуды, частоты или фазы сигнала). При переработке информации в полиграфии применяются в основном сигналы двух видов: электрические и световые. Этап отображения информации обычно предусматривает использование оптической формы сигнала.

На этапе хранения информация определяется состоянием хранителя (носителя) информации.

В зависимости от объекта, о котором собирается информация, она делится на технологическую, техническую, научную, экономическую, социальную, медицинскую, культурную.

16.1. Количественная мера информации

В теории информации понятие информации в чистом виде не определяется. Необходимым и достаточным для построения теории является понятие количества информации. Вводимая мера информа-

ции должна быть удобной для анализа и синтеза, для передачи и хранения информации и «нечувствительной» к смыслу, ценности и степени правдивости информации.

Количество информации должно определяться через нечто общее. Этим общим, характеризующим факт получения произвольной информации, является, во-первых, наличие опыта. Всякая информация получается только в результате опыта. Опытом может быть чтение книги, прослушивание радио, визуальное наблюдение, измерение некоторого параметра тем или иным прибором и т. д. Во-вторых, до испытания должна существовать некоторая неопределенность в том или ином исходе. В самом деле, если бы получателю до опыта было известно, какое сообщение он получит, то, получив его, он не приобрел бы никакого количества информации.

Таким образом, до опыта всегда имеется большая или меньшая неопределенность в интересующей нас ситуации. Разность между количествами неопределенности до и после опыта можно отождествить с количеством полученной информации в результате такого испытания.

В этой ситуации к количеству информации (или, что то же самое, к количеству неопределенности до опыта) можно предъявить три априорных условия:

1. Количество получаемой информации больше в том опыте, у которого большее число возможных исходов.

2. Опыт с единственным исходом несет количество информации, равное нулю.

3. Количество информации от двух независимых опытов должно равняться сумме количеств информации от каждого из них.

Функцией от n, удовлетворяющей трем поставленным условиям, является логарифмическая функция. Итак, количество информации от опыта с n исходами при условии, что после опыта неопределенность отсутствует, равно

$$l = c \log_{\alpha} n ,$$

где c и α – произвольные постоянные.

Если исходы опыта считать равновероятными, т. е. вероятность любого исхода равна

$$P = \frac{1}{n}, \mathrm{P} = 1/\mathrm{n}$$

то

$$l = -c \log_{\alpha} P .$$

Принимая $c = 1$, $\alpha = 2$, вычислим количество информации в единицах бит, получаемых в результате опыта с двумя равновероятными исходами. Другая единица («нат») получается, если использовать натуральные логарифмы, обычно она употребляется для непрерывных величин.

Если исходы опыта неравновероятны, то усредненное количество информации будет

$$l = 1\sum OP,$$

$$c \log_2 P_i = H_i,$$

где H_i – энтропия.

Это означает, что каждой реализации результатов наблюдений соответствует своя энтропия. Эта энтропия является величиной случайной и априорной, после проведения опыта она будет равна нулю. Поясним это следующим примером. Известна вероятность, что событие произойдет $P_i = 7/8$, и вероятность, что оно не произойдет $P_i = 1/8$, т. е. вероятности показывают с большой степенью достоверности, что событие произойдет. Если нам сообщают, что событие произошло, то количество информации этого сообщения будет

$$l_1 = H_1 = \log_2 8 - \log_2 7 = 0{,}193 \text{ бит.}$$

Если мы получаем сообщение, что событие не произошло, то количество информации

$$l_2 = H_2 = \log_2 8 = 3 \text{ бит.}$$

Если априори известно, что вероятность события равна $P = 1$, то сообщение об этом событии не дает нам никакой информации.

16.2. Энтропия как мера неопределенности

Энтропия – мера вариантности системы или мера неопределенности результатов наблюдения какого-либо события.

Результаты случайных событий нельзя узнать заблаговременно, так как им присуща неопределенность. Отсюда вытекает необходимость введения количественной меры неопределенности наблюдения случайных событий.

Формула, определяющая энтропию, выводится через результаты наблюдений, которые нумеруются в двоичной системе счисления с учетом следующих правил:

1. Равновероятные результаты наблюдений обозначаются одним и тем же количеством двоичных знаков.

2. Чем больше вероятность результата, тем меньшим числом двоичных знаков он нумеруется.

Результаты наблюдений разбиваются на две группы так, чтобы сумма вероятностей в каждой группе была близка к 1/2, но при этом так, чтобы в первой группе были события с большими вероятностями. Всем результатам первой группы приписывается первый двоичный знак 1, а второй – 0. Чтобы определить второй двоичный знак нумерации результатов, каждая из двух групп разбивается еще на две подгруппы. Сумма вероятностей в этих подгруппах должна быть примерно равной 1/4. Первой и третьей подгруппам присваивается второй двоичный знак 1, а второй и четвертой – 0. Продолжив такое разбиение на все более мелкие подгруппы и обозначив номер результата наблюдений m-значным двоичным числом, определим вероятности всех возможных испытаний по формуле

$$P_1 = 2^{-mi},$$

где $i = 1, 2, ..., m$, а $m_1, m_2, \ldots, m_m$ – целые положительные числа.

Количество двоичных знаков представляет собой случайную величину, вероятности значений которой соответственно:

$$P = \sum_{i=1}^{m} 2^{-mi} = 1.$$

За меру неопределенности результатов наблюдений целесообразно принять математическое ожидание числа двоичных знаков:

$$H = M\left[m_i\right] = m_1 2^{-m_1} + m_2 2^{-m_2} + ... + m_m 2^{-m_m}. \qquad (16.1)$$

Так как $P_i = 2^i$, то $m_i = \log_2 P_i$. Поэтому формула может быть переписана в следующем виде:

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i .$$

Эта формула была предложена Шенноном для количественного определения энтропии. Она является обобщением меры неопределенности результатов наблюдений, предложенной ранее Хартли для случая равновероятных наблюдений. Единицей измерения энтропии является один двоичный знак бит. Если же в формуле (16.1) был бы натуральный логарифм $\ln P_i$, то единицей измерения был бы

нат. Необходимо отметить, что эта натуральная единица используется очень редко.

Вполне определенный смысл имеет и следующая зависимость:

$$H_i = -\log_2 P_i.$$

Она означает, что каждой реализации результатов наблюдений соответствует своя энтропия. Информация, получаемая после выяснения реализации, будет выражаться:

$$l_i = H_i = -\log_2 P_i.$$

16.3. Свойства энтропии и их доказательства

Свойство 1. Энтропия является вещественной неотрицательной величиной. Так как P_i изменяется от 0 до 1, то $\log_2 P_i$ отрицателен и, следовательно, величина $P_i \log_2 P_i$ является положительной.

Свойство 2. Энтропия – величина ограниченная. Для слагаемых в диапазоне изменения P_i от 0 до 1 ограниченность энтропии очевидна. Найдем теперь предел, к которому стремится произвольное слагаемое – $P_i \log_2 P_i$ при $P_i \to 0$, так как в этом случае величина $\log_2 P_i$ неограниченно возрастает:

$$\lim_{P_i \to 0} (-P_i \log_2 P_i) = \lim_{P_i \to 0} \frac{\log_2 \frac{1}{P_i}}{\frac{1}{P_i}}.$$

Введем обозначение $\frac{1}{P_i} = x$, тогда:

$$\lim_{P_i \to 0} (-P_i \log_2 P_i) = \lim_{x \to \infty} \frac{\log_2 x}{x}.$$

Используя правило Лопиталя, находим

$$\lim_{P_i \to 0} (-P_i \log_2 P_i) = \lim_{x \to \infty} \frac{\frac{1}{x} \log_2 e}{1} = 0.$$

Таким образом, и в этом случае энтропия является ограниченной.

Свойство 3. Энтропия обращается в нуль в том случае, когда вероятность одного из состояний равна единице. Полагая $P_1 = 1$,

$P_2 = P_3 = \ldots = P_m = 0$, из формулы находим $H = 0$. Это означает, что состояния событий полностью определены:

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i.$$

Свойство 4. Энтропия максимальна, когда все события равновероятны. Чтобы доказать это свойство, необходимо воспользоваться методом множителей Лагранжа. Введем некоторую функцию m и, производя определенные математические преобразования, получим

$$H = -\sum_{i=1}^{m} \frac{1}{m} \log_2 \frac{1}{m} = -\log_2 m.$$

Именно в таком виде Р. Хартли получил формулу для энтропии при равной вероятности событий.

Свойство 5. Энтропия двух независимых опытов с числом исходов, равных m_1 и m_2, является суммой энтропии каждого из опытов. Совместная энтропия определяется формулой

$$H_{1,2} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} P(i, j) \log_2 P(i, j),$$

где $P(i,j) = P_i P_j$, а i и j – вероятности i-го и j-го исхода соответственно. Формула может быть переписана в следующем виде:

$$H_{1,2} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} P_i P_j (\log_2 P_i - \log_2 P_j).$$

Преобразовав эту сумму, имеем

$$H_{1,2} = \sum_{i=1}^{m_1} P_i \log_2 P_i \sum_{j=1}^{m_2} P_j - \sum_{i=1}^{m_2} P_j \log_2 P_j \sum_{j=1}^{m_1} P_i .$$

Учитывая очевидные равенства

$$\sum_{i=1}^{m_1} P_i = 1 \text{ и } \sum_{j=1}^{m_2} P_j = 1,$$

получим

$$H_{1,2} = H_1 + H_2.$$

Таким образом, свойство 5 также доказано.

16.4. Энтропия взаимосвязанных событий и их свойства

Пусть имеется некоторый источник сообщений, связанный каналом передачи с приемником информации. Взаимосвязь задается условной вероятностью $P(j|i)$ того, что после появления символа j появляется символ L. Неопределенность всех символов по отношению к символу i характеризуется энтропией

$$H_i = -\sum_{j=1}^{m} P(j \mid i) \log_2 P(j \mid i).$$

Общая энтропия выражается математическим определением $H(j|i)$ как случайной величины энтропии и означает, что после появления символа j появится символ i:

$$H(j \mid i) = M\left[H_i\right] = P_1 H_1 + P_2 H_2 + \ldots + P_m H_m.$$

Объединяя эту схему и подставляя значение H_i, получаем

$$H(j \mid i) = -\sum_{i=1}^{m1} P_i \sum_{j=1}^{m2} P(j \mid i) \log_2 P(j \mid i). \qquad (16.2)$$

Условная энтропия записывается в виде $H_x(y)$, что выражает неопределенность значения y при заданных значениях x.

Рассмотрим теперь два параллельно работающих источника, выдающих дискретные сигналы x и y. Появление на выходе обоих источников любой пары значений сигналов x_i и y_i определяется вероятностью $P(i,j)$. Если число возможных значений каждого из сигналов равно m, то общее число всех возможных комбинаций сигналов будет равно m^2.

Систему двух сигналов x и y можно заменить одним источником, выдающим сигнал z, число возможных значений которого $s = m^2$, а вероятность появления комбинаций $P_к = P(i, j)$. Энтропия этого источника будет

$$H(z) = -\sum_{k=1}^{5} P_k \log_2 P_k.$$

Учитывая, что $H(z) = H(x, y)$, находим совместную энтропию двух параллельно работающих и связанных друг с другом источников:

$$H(x \mid y) = -\sum_{i=1}^{m} \sum_{j=1}^{m} P(j \mid i) \log_2 P(j \mid i). \qquad (16.3)$$

Найдем зависимость между совместными и условными энтропиями. Условная вероятность $P(i|j)$ – вероятность того, что появление сигнала x на выходе первого источника сопровождается появлением сигнала y на выходе второго, – равна

$$P(j \mid i) = \frac{P(j,i)}{P(i)},$$

а условная энтропия

$$H_x(y) = -\sum_{i=1}^{m} P_i \sum_{j=1}^{m} P(j \mid i) \log_2 P(j \mid i). \qquad (16.4)$$

Очевидные соотношения

$$H(x) = -\sum_{i=1}^{m} P_i \log_2 P_i,$$

$$H(y) = -\sum_{j=1}^{m} P_j \log_2 P_j$$

можно преобразовать к следующей форме:

$$\begin{aligned} H(x) &= -\sum_{i=1}^{m} \sum_{j=1}^{m} P(j \mid i) \log_2 P_i, \\ H(y) &= -\sum_{i=1}^{m} \sum_{j=1}^{m} P(j \mid i) \log_2 P_j. \end{aligned} \qquad (16.5)$$

Вычтем из формулы (16.2) формулу (16.4), тогда имеем

$$H(x,y) - H(x) = -\sum_{i=1}^{m} \sum_{j=1}^{m} P(j \mid i) \log_2 P(i,j) + \sum_{i=1}^{m} \sum_{j=1}^{m} P(j \mid i) \log_2 P_i =$$

$$= -\sum_{i=1}^{m} \sum_{j=1}^{m} P(j \mid i) \log_2 \frac{P(i,j)}{P_i}.$$

Подставляя сюда значения вероятностей из формулы (16.3), получаем:

$$H(x,y) - H(x) = -\sum_{i=1}^{m} P_i \sum_{j=1}^{m} P(j \mid i) \log_2 P(j \mid i).$$

Правая часть этой зависимости равна $H_x(y)$, поэтому

$$H(x,y) - H(x) = H_x(y).$$

Аналогичным путем находим:

$$H(x,y) - H(y) = H_y(x).$$

16.5. Свойства энтропии взаимосвязанных событий

Свойство 1. Сумма энтропии двух источников сигналов x и y равна или больше их совместной энтропии:

$$H(x,y) \le H(x) + H(y).$$

В частном случае независимых сигналов x и y и $P(i,j) = P_iP_j$, как это было показано ранее, имеет место равенство

$$H(x,y) = H(x) + H(y).$$

Свойство 2. Энтропия источника со взаимосвязанными сигналами меньше энтропии того же источника с независимым появлением сигналов:

$$H(x) \ge H_y(x),$$
$$H(y) \ge H_x(y).$$

16.6. Количество информации в дискретных сообщениях

Если источник сигнала выдает m различных сигналов, из которых формируются сигналы в n момент считывания, то общее число возможных реализаций сигнала N будет равно

$$N = m^n.$$

Втом случае, когда n велико и хотя бы две вероятности появления сигналов $P_1, P_2, \ldots, P_m$ различны, т. е. не равновероятны, то все реализации сигнала от источника могут быть разбиты на две группы – высоковероятную и маловероятную.

При оценке количества информации в сигнале, вырабатываемом источником, целесообразно и достаточно учитывать только высоковероятную часть реализаций, так как вклад маловероятной части в количественную оценку информации будет незначителен. Обозначив количество таких реализаций числом N_1, можем записать количество информации как

$$N_1 = \frac{n!}{n_1!n_2!...n_m!},$$

где $n_1, n_2, \ldots, n_m$ – число сигналов в каждой из реализаций.

Количество информации равно логарифму с основанием два:

$$I = \log_2 N_1.$$

После подстановки значения числа N_1 и некоторых преобразований получим окончательную формулу количества информации

$$I = -n\sum_{i=1}^{m} P_i \log_2 P_i .$$

Сравнивая между собой формулу определения энтропии и вышеприведенную формулу, можно установить, что

$$I = nH.$$

где n – число сигналов в каждой реализации.

Отсюда следует совпадение энтропии с количеством информации на один отсчет сигнала.

Если исходы неравновероятны, то усредненное количество информации $I = H_i$.

16.7. Количество информации во взаимосвязанных объектах

Если рассматривать два объекта, то количество информации в источнике коррелированного сигнала за один отсчет будет равно энтропии $H(j|i)$. За n отсчетов количество информации будет

$$I_{\text{кор}} = nH(j \mid i).$$

Подставляя сюда значение энтропии, находим

$$I_{\text{кор}} = -n\sum_{i=1}^{m} P_i \sum_{i=1}^{m} P(j \mid i) \log_2 P(j \mid i).$$

Когда источники двух взаимосвязанных сигналов x и y работают параллельно и совместная энтропия двух источников определена $H(x,y)$, то количество информации в них будет

$$I_{CB} = nH(x, y).$$

После подстановки сюда энтропии находим

$$I_{CB} = -n\sum_{i=1}^{m} P_i \sum_{i=1}^{m} P(j \mid i) \log_2 P(j \mid i).$$

Рассмотрим задачу о количестве информации при приеме сигналов в условиях случайных помех. Значение принятого сигнала y_j в этих условиях за счет искажений перестанет полностью соответствовать переданному сигналу x_i.

Условная энтропия $H_y(x)$ при этом будет характеризовать степень неопределенности сигнала x при приеме конкретного сигнала y.

Количество информации при этом будет равно разности

$$I = n[H(x) - H_y(x)].$$

После подстановки сюда значений энтропии и преобразований получаем

$$I = -n\sum_{i=1}^{m}\sum_{i=1}^{m} P(j \mid i)\log_2\left[\frac{P(j \mid i)}{P_i}\right].$$

16.8. Количество информации в непрерывных сигналах

Для определения количества информации в одиночном сигнале необходимо воспользоваться зависимостью непрерывного квантованного сигнала:

$$I = n[h(x) - \log_2 \delta_x].$$

Запишем значения количества информации для различных законов распределений случайных величин:

– нормального:

$$I = n\log_2\left(\sqrt{2\pi e}\,\frac{\sigma_x}{\delta_x}\right);$$

– экспоненциального:

$$I = n\log_2\left(\frac{ae}{\delta_x}\right);$$

– равномерного:

$$I = n\log_2\left(\frac{2\alpha}{\delta_x}\right).$$

Отсюда следует, что количество информации возрастает с уменьшением шага квантования сигнала по амплитуде δ_x и увеличением дисперсии.

В том случае, когда два источника выдают сигналы параллельно, количество информации в общем виде в них будет

$$I_{CB} = n[h(x,y) - \log_2 \delta_x \delta_y].$$

Приведем значение количества информации для нормального распределения:

$$I_{\text{СВ}} = n \log_2 \left[2\pi e \frac{\sigma_x \sigma_y}{\sigma_x \sigma_y} \sqrt{1 - r^2} \right].$$

Из этой формулы также усматривается зависимость количества информации от дисперсий сигналов и шагов их квантования.

Лекция 17. ПЕРЕДАЧА ИНФОРМАЦИИ ПО КАНАЛАМ СВЯЗИ

17.1. Классификация каналов передачи информации

Совокупность операторов, технических средств передачи и приема сигналов, физических сред и средств связи образует канал передачи информации. Каналы передачи информации могут быть дискретными и непрерывными.

Информационная модель канала задается совокупностью символов на его входе и выходе и описанием вероятностных свойств передачи отдельных символов. Канал может иметь множество состояний и переходить из одного состояния в другое как с течением времени, так и в зависимости от последовательности передаваемых сигналов.

Состояние канала характеризуется условной вероятностью $P(x_i|y_i)$ того, что передаваемый символ x_i будет воспринят на выходе как символ y_i. Значения этих вероятностей зависят от многих факторов: свойств сигналов, метода кодирования, наличия случайных помех в канале передачи, принципа декодирования, психофизиологического состояния оператора.

Если эти переходные вероятности не зависят от времени, то канал называется стационарным, или иначе информационным каналом с установившейся связью. В нестационарном канале связи переходные, вероятности зависят от времени. Если они зависят от его предыдущего состояния, то канал носит наименование канала с памятью. В противном случае его называют каналом без памяти.

При передаче двух символов 0 и 1 стационарный канал без памяти определяется четырьмя условными переходными вероятностями: $P(1|1)$, $P(1|0)$, $P(0|1)$, $P(0|0)$, где $P(0|0)$ и $P(1|1)$ – вероятности неискаженной передачи символов 0 и 1, а $P(1|0)$ и $P(0|1)$ – вероятности искажения символов 0 и 1. Когда вероятности искажения символов $P(1|0)$ и $P(0|1)$ равны друг другу, канал носит наименование двоичного симметричного канала.

17.2. Передача информации при отсутствии помех

При отсутствии помех в канале связи каждому сообщению на входе канала соответствует вполне определенное сообщение на его выходе.

Количество информации в сообщении связано с энтропией известной зависимостью:

$$I = nH, \tag{17.1}$$

где n – число отсчетов.

Если T – длительность сообщений, а Δt – время на один отсчет, то

$$n = \frac{T}{\Delta t}.$$

Подставив это значение в формулу (17.1) и разделив левую и правую части на длительность сообщений, получаем

$$\frac{I}{T} = \frac{H}{\Delta t} = V. \tag{17.2}$$

Величина V – скорость создания информации источником, ее размерность – бит в секунду. Связь между интервалом отсчетов Δt и полосой частот сигнала была сформулирована теоремой В. А. Котельникова. В соответствии с этой теоремой

$$T_{\text{И}} = \frac{1}{2\Delta f},$$

$$\Delta t = \frac{1}{2\Delta f},$$

где Δf – полоса частот, занимаемых сигналом. Подставляя это значение в формулу (17.2), имеем:

$$V = 2H\Delta f.$$

Рассмотрим некоторую последовательность сообщений длительностью каждое t_1, t_2, …, t_m. Если общая длительность передачи всех сообщений T, а число таких сообщений $N(T)$, то количество информации, проходящей через канал, будет

$$I_{\text{К}} = \log_2 N(T).$$

Скорость же передачи информации по каналу

$$V_{\text{К}} = \frac{I_{\text{К}}}{T} = \frac{\log_2 N(T)}{T}.$$

Из этой формулы следует, что скорость передачи информации зависит от длительности передачи T.

Пропускная способность канала – максимальная скорость передачи, которая возможна для данного канала. Она определяется следующим пределом:

$$V_{\text{пр}} = \lim_{T \to \infty} \frac{\log_2 N(T)}{T}.$$

Если скорость создания сообщений источников равна V, то с этой же скоростью информация должна передаваться по каналу $V_{\text{К}} = V$. Тогда из формул (17.1) и (17.2) имеем

$$V = \frac{nH}{T}.$$

Обозначив скорость передачи информации как V пропускания ($V_{\text{пр}}$), можем записать следующую пропорцию:

$$\frac{n}{T} = \frac{V_{\text{пр}}}{H},$$

где отношение n / T будет представлять максимальное количество элементов, передаваемое в секунду по каналу.

Максимальное число элементов можно передавать по каналу только при оптимальном кодировании, при котором скорость передачи сообщений по каналу равна его пропускной способности. В реальных каналах скорость передачи информации меньше пропускной способности канала.

Значения пропускной способности различных технических и биологических каналов передачи информации приведены в таблице.

Виды каналов связи	$V_{\text{пр}}$, десятичных единиц информации
Технические:	
телевизионные,	миллионы – десятки миллионов
телефонные,	тысячи – десятки тысяч
фототелеграфные, телеграфные	десятки – сотни
Биологические:	
органы:	миллионы
зрения,	тысячи
слуха,	десятки тысяч
осязания,	единицы – десятки
обоняния,	единицы
вкуса,	единицы
центральная нервная система	

Пропускная способность операторов характеризуется следующими данными: корректорское чтение – 18 бит/с, чтение вслух – 30 бит/с,

чтение про себя – 45 бит/с, ввод данных в ЭВМ с контролем по экрану – 2500–3500 знак/ч.

Представляет интерес нахождение зависимости между оптимальными длительностями t_i символов x_i, посредством которых кодируется сообщение, и вероятностями появления этих символов P_i. Эту задачу оптимизации длительности символов рассмотрим из условия максимальной скорости передачи информации.

Если проходящий по каналу сигнал содержит n_1 символов x_1, n_2 символов x_2 и n_m символов x_m, то общее число элементов сигнала равно

$$m = \sum_{j=1}^{m} n_j,$$

а частота их появления будет определятся вероятностью

$$P_j = \frac{n_j}{n}.$$

При этом должно соблюдаться условие

$$\sum_{j=1}^{m} P_j = 1. \qquad (17.3)$$

Длительность сигнала определяется следующей суммой:

$$T = \sum n_j t_j$$

или, если значения n_j выразить через вероятность $n_j = P_j n$,

$$T = n \sum P_j t_j. \qquad (17.4)$$

Количество информации, проходящее по каналу, определяется как

$$I = \log_2 N(T) = -n \sum_{j=1}^{m} P_j \log P_j.$$

После максимизации этой функции, выполнения условий (17.3) и (17.4) и целого ряда преобразований находим значение вероятности кодовых символов:

$$P_i = 2^{-\lambda t_i}.$$

Такая экспоненциальная зависимость между длительностями кодовых символов и их вероятностями позволяет обеспечить максимальную скорость передачи информации по каналу. Пропускная способность канала в этом случае будет

$$V_{\text{пр}} = \frac{-n\sum_{j=1}^{m} P_j \log_2 P_j}{T} \text{ или } V_{\text{пр}} = \lambda_2 \frac{n\sum_{j=1}^{m} P_j t_j}{T}.$$

Но так как на основании формулы (17.4)

$$T = n\sum P_j t_j ,$$

то пропускная способность канала равна

$$V_{\text{пр}} = \lambda_2. \qquad (17.5)$$

Из полученной зависимости и формулы (17.5) можно получить простую зависимость для определения длительностей кодовых символов:

$$t_j = \frac{\log_2 P_j}{V_{\text{пр}}}. \qquad (17.6)$$

При кодировании сигналов двоичным кодом кодовые комбинации имеют фиксированные (дискретные) длины, поэтому точное согласование вероятностей P_i и длительностей кодовых символов t_i невозможно. Полученная формула (17.6) позволяет только выбрать двоичный код, достаточно близкий к оптимальному.

17.3. Передача информации при наличии помех

Помехи, действующие в канале передачи, вызывают искажение полезного сигнала, что приводит к потере некоторой части передаваемой информации. Пропускной способностью канала в этом случае считается максимальная скорость передачи информации в условиях заданного уровня помех, при этом вероятность ошибки передачи является сколь угодно малой.

Если бы помехи отсутствовали, то передаваемый сигнал x и принимаемый сигнал y были бы одинаковыми. Наличие помех приводит к неопределенности значения сигнала x при конкретном сигнале y.

Если по некоторому каналу при отсутствии помех передается информация, равная I_K / T, то при их наличии эта информация уменьшается. Ее уменьшение будет пропорционально условной энтропии $H_y(x)$, тогда скорость, с которой полезная информация передается по каналу, определяется

$$V_y = \frac{I_{\text{К}}}{T} - \frac{nH_y(x)}{T}.$$

При вычислении условной энтропии следует обратить внимание на возможность различного объема алфавитов входного и выходного сигналов. Поэтому если объем алфавита сигнала x равен m, а сигнала y – l, то условная энтропия будет

$$H_y(x) = -\sum_{i=1}^{m}\sum_{j=1}^{l} P(j,\ i)\log_2 P(j\,|\,i).$$

Так как $I_К = nH(x)$, то скорость передачи информации равна

$$V_y = \frac{n}{T}[H(x) - H(y)].$$

Пропускная способность канала будет равна максимальной скорости передачи информации, следовательно:

$$V_{пр} = \frac{n}{T}[H(x) - H_y(x)]_{\max}.$$

Пропускная способность двоичного симметричного канала будет

$$V_y = V_T[\max H(y) + P\log_2 P + (1-P)\log_2(1-P)].$$

Но максимум $H(y)$ достигается при равновероятном появлении символов, причем $\max H(y) = 1$, поэтому

$$V_y = V_T[1 + P\log_2 P + (1-P)\log_2(1-P)].$$

Анализ этой зависимости показывает, что при изменении вероятности P от 0 до 1/2 пропускная способность изменяется от 1 до 0. Если $P = 0$, т. е. помеха в канале передачи отсутствует, то его пропускная способность равна 1. При большом уровне шума, когда $P = 0{,}5$, использование такого канала невозможно, так как в приемнике с равным успехом можно принимать либо значение 0, либо 1. Пропускная способность в этом случае равна нулю.

Передача информации непрерывными квантованными сигналами может осуществляться как при отсутствии, так и при наличии помех в канале передач. При этом пропускная способность зависит не только от характеристик полезного сигнала и помехи, но и от параметров канала передачи информации. Зависимость пропускной способности канала от дисперсии помехи и полезного сигнала, полученная Шенноном, имеет вид

$$V_{пр} = \Delta f \log_2\left(1 + \frac{D_x}{D_y}\right), \qquad (17.7)$$

где Δf – полоса пропускания частот каналом, Гц; D_x и D_y – дисперсии полезного сигнала и помехи.

Эта формула может быть использована и для расчета пропускной способности в случае отсутствия помех в канале передачи. Для этого в формулу подставляют дисперсию ошибки квантования сигнала. Если квантование сигнала осуществляется по уровню при равновероятном появлении символов, то дисперсия полезного сигнала будет

$$D_x = \frac{x_m^{\,2}}{3},$$

где x_m – максимальное значение сигнала.

Дисперсия же помехи при равном квантовании имеет вид

$$D_y = \frac{\delta_x^{\,2}}{12}.$$

Подставляя значения дисперсий в выражение (17.7), получаем

$$V_{\text{пр}} = \Delta f \log_2\left(1 + \frac{4x_m^{\,2}}{\delta_x^{\,2}}\right).$$

Так как

$$\frac{4x_m^{\,2}}{\delta_x^{\,2}} = m^2,$$

где m – число уровней квантования сигнала, то получим

$$V_{\text{пр}} = \Delta f \log_2\left(1 + m^2\right).$$

Так как $m^2 \gg 1$, то пропускная способность будет

$$V_{\text{пр}} = \Delta f \log_2 m^2.$$

Вероятность ошибки при передаче символа в случае амплитудно-импульсной модуляции представляет собой вероятность того, что уровень помехи превышает половину шага квантования сигнала δ_x.

Если уровень помехи превысит половину шага квантования, то принятый символ уже может быть отнесен к соседнему символу.

Вычисленные по вышеприведенной формуле вероятности в зависимости от коэффициента K, приведены в таблице, где

$$K = \frac{\delta_x}{\sqrt{D_y}}$$

K	7	8	9	10	11
P	$5 \cdot 10^{-4}$	$6{,}6 \cdot 10^{-5}$	$7 \cdot 10^{-6}$	$6 \cdot 10^{-7}$	$4 \cdot 10^{-8}$

Уровень квантования m для случая амплитудно-импульсной модуляции определяется

$$m^2 = \frac{3D_x}{\Delta f k S_y},$$

где $S_y(\omega)$ – спектральная плотность белого шума в канале передач, определяемая как

$$S_y = \frac{D_y}{\Delta f}.$$

Из всего вышесказанного можно сделать вывод, что скорость передачи информации по каналу можно повысить за счет уменьшения дисперсии помехи, увеличения шага квантования, амплитуды сигнала, полосы пропускания частот каналом, повышения допускаемой вероятности ошибки передачи информации.

ЛИТЕРАТУРА

1. Долгова Т. А. Методы моделирования полиграфических процессов: учеб. пособие для студентов высших учебных заведений по полиграфическим специальностям / Т. А. Долгова. – Минск: БГТУ, 2009. – 166 с.

2. Ефимов М. В. Теоретические основы переработки информации в полиграфии. В 2 кн. Кн. 1 : учеб. для вузов / М. В. Ефимов. – М.: МГУП, 2001. – 340 с.

3. Гасов В. М. Информационные технологии в издательском деле и полиграфии: в 2 кн. / В. М. Гасов, А. М. Цыганенко. – М.: МГУП «Мир книги», 1999.

4. Одинокова Е. В. Проектирование полиграфических машин: учеб. для вузов / Е. В. Одинокова, Г. Б. Куликов, И. Ш. Герценштейн. – М.: МГУП, 2003. – 411 с.

ОГЛАВЛЕНИЕ

Учебное издание

Барташевич Святослав Александрович

МОДЕЛИРОВАНИЕ СИСТЕМ ОБРАБОТКИ ИНФОРМАЦИИ

Электронный курс лекций

Редактор *А. Д. Микитюк*
Компьютерная верстка *А. Д. Микитюк*
Корректор *А. Д. Микитюк*

Издатель:
УО «Белорусский государственный технологический университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/227 от 20.03.2014.
Ул. Свердлова, 13а, 220006, г. Минск.