УДК 004.6+51
ББК 32.973.26-018.2
С40

Сирота А. А.

С40 Методы и алгоритмы анализа данных и их моделирование в MATLAB: учеб. пособие. — СПб.: БХВ-Петербург, 2016. — 384 с.: ил. — (Учебное пособие)

ISBN 978-5-9775-3778-0

Рассматриваются модели, методы и алгоритмы анализа данных, используемые в современных системах обработки информации. Приводятся основные понятия и определения общей теории информационных систем, анализируется типовая структура систем извлечения информации и систем обработки информации, рассматриваются типовые задачи анализа данных в системах обработки информации и базовые подходы для их решения. Представлены методы и алгоритмы, используемые при решении задач оценивания, регрессии и фильтрации, распознавания и кластеризации образов. Рассматриваются классические и современные реализации указанных алгоритмов в рамках статистического и детерминистского подходов. В книге и на сайте издательства приводятся примеры построения компьютерных моделей в среде MATLAB, представляющих программную реализацию алгоритмов анализа данных.

Для студентов компьютерных направлений и специальностей

УДК 004.6+51
ББК 32.973.26-018.2

Р е ц е н з е н т ы:

А. Г. Буховец, д-р техн. наук, профессор кафедры прикладной математики и математических методов в экономике Воронежского государственного аграрного университета имени Императора Петра I

Е. А. Самойлин, д-р техн. наук, профессор кафедры Военного учебно-научного центра Военно-воздушных сил «Военно-воздушная академия имени профессора Н. Е. Жуковского и Ю. А. Гагарина»

Группа подготовки издания:

Главный редактор	*Екатерина Кондукова*
Зам. главного редактора	*Людмила Еремеевская*
Зав. редакцией	*Екатерина Капалыгина*
Редактор	*Григорий Добин*
Компьютерная верстка	*Ольги Сергиенко*
Корректор	*Зинаида Дмитриева*
Дизайн серии	*Инны Тачиной*
Оформление обложки	*Марины Дамбиевой*

Подписано в печать 31.08.16.
Формат 70×100 1/16. Печать офсетная. Усл. печ. л. 30,96.
Тираж 700 экз. Заказ №
"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

Первая Академическая типография "Наука"
199034, Санкт-Петербург, 9 линия, 12/28

ISBN 978-5-9775-3778-0

Оглавление

Введение 9

Глава 1. Информационные процессы и системы. Обработка информации, анализ данных, машинное обучение 13
1.1. Основные понятия и определения. Математическое описание систем в рамках теоретико-множественного подхода 13
1.2. Классификация систем. Информационные системы, информационные процессы, информационные технологии 19
1.3. Задачи анализа данных в системах обработки информации и базовые подходы для их решения 30

ЧАСТЬ I. ОЦЕНИВАНИЕ, РЕГРЕССИЯ, ФИЛЬТРАЦИЯ 39

Глава 2. Математические описания и моделирование случайных величин и случайных векторов 41
2.1. Математические описания и выборочные характеристики случайных величин и случайных векторов 41
2.1.1. Квадратичные формы и линейные преобразования случайных векторов 50
2.1.2. Выборочные характеристики случайных величин и случайных векторов 51
2.2. Моделирование случайных величин и случайных векторов 53
2.2.1. Моделирование простейших случайных величин на основе стандартных датчиков случайных чисел 54
2.2.2. Моделирование случайных величин с произвольными законами распределения 57
Метод нелинейного функционального преобразования 57
Метод исключений (метод фон Неймана) 58
2.2.3. Моделирование случайных векторов с заданной матрицей ковариаций 62

Глава 3. Основы теории оценивания и регрессионного анализа данных 69
3.1. Общая характеристика задач оценивания. Оценка параметров в рамках статистического и детерминистского подходов 69
3.1.1. Статистический подход к решению задачи параметрического оценивания. Методы максимума правдоподобия и максимума апостериорной вероятности 69
Метод максимума правдоподобия 70

Байесовское оценивание 73
Моделирование алгоритмов 76
Теорема о нормальной корреляции 79
3.1.2. Детерминистский подход к решению задачи параметрического оценивания. Метод наименьших квадратов 80
Линейный случай....82
Нелинейный случай....83
Моделирование алгоритма 85
3.2. Непараметрическая оценка плотностей распределения вероятностей 89
3.2.1. Оценка плотности распределения вероятностей на основе метода Парзена....89
Моделирование алгоритма 92
3.2.2. Оценка плотности распределения вероятностей на основе метода k-ближайших соседей 102
Моделирование алгоритма 103
3.2.3. Нелокальные методы оценивания плотности распределения вероятностей 111
Гистограммный метод оценивания....111
Метод аппроксимации с использованием ортогональных функций....112
3.3. Основы регрессионного анализа данных....113
3.3.1. Постановка и решение задачи регрессии в рамках статистического подхода 114
3.3.2. Постановка и решение задачи регрессии в рамках детерминистского подхода по методу наименьших квадратов 116
Линейная параметрическая регрессия....117
Проверка значимости модели регрессии....119
Моделирование алгоритма 121
3.3.3. Метод псевдообратной матрицы и метод регуляризации в задачах регрессии....124
Моделирование алгоритма 125
3.3.4. Расширения линейной регрессии и нелинейная регрессия 128
Моделирование алгоритма 129

Глава 4. Фильтрация — оценивание изменяющихся параметров состояния объектов....133
4.1. Основные положения теории оптимальной марковской фильтрации в дискретном времени....133
4.1.1. Общая методика решения задач оптимальной фильтрации в дискретном времени 133
4.1.2. Постановка и решение задачи оптимальной линейной фильтрации....135
Моделирование алгоритмов 138
4.2. Расширения задачи оптимальной линейной фильтрации....145
4.2.1. Негауссовские модели параметров и оптимальный в классе линейных фильтр 145
4.2.2. Расширенный фильтр Калмана....150
Моделирование алгоритмов 150
4.2.3. Адаптивная постановка задачи фильтрации и метод разделения....157
Моделирование алгоритма 160
4.3. Синтез и анализ алгоритмов фильтрации для оценки состояния объектов в условиях аномальных наблюдений....164
4.3.1. Модели получения аномальных наблюдений....164

4.3.2. Синтез и анализ различных типов алгоритмов фильтрации в условиях аномальных наблюдений 167
Оптимальный в классе линейных фильтр 167
Моделирование алгоритма 170
Условно линейный фильтр 173
Моделирование алгоритма 176
Оптимальный нелинейный фильтр 181

ЧАСТЬ II. РАСПОЗНАВАНИЕ И КЛАСТЕРИЗАЦИЯ 193

Глава 5. Основы статистической теории распознавания образов 195
5.1. Байесовская теория принятия решения применительно к задаче распознавания образов 195
5.1.1. Синтез решающих правил на основе различных критериев оптимальности 196
Критерий минимума условного риска 196
Критерии максимума апостериорной вероятности и максимального правдоподобия 198
Обобщенная структура решающего правила. Понятие разделяющей функции 199
5.1.2. Анализ решающих правил. Способы определения вероятностей ошибок распознавания 201
5.2. Распознавание образов, описываемых гауссовскими случайными векторами 208
5.2.1. Распознавание образов, описываемых гауссовскими случайными векторами с одинаковыми матрицами ковариаций 209
Моделирование алгоритма 212
5.2.2. Распознавание образов, описываемых гауссовскими случайными векторами с различными матрицами ковариаций 218
Моделирование алгоритма 219
5.3. Распознавание образов, описываемых произвольными законами распределения 224
5.3.1. Распознавание образов в предположении статистической независимости признаков 224
5.3.2. Распознавание образов в случае статистически независимых дискретных признаков 226
5.3.3. Распознавание на основе бинарных признаков (на примере анализа бинарных изображений) 228
Моделирование алгоритма 230
5.4. Распознавание образов в условиях параметрической и непараметрической неопределенности на основе обучения с учителем 233
5.4.1. Распознавание образов в условиях параметрической неопределенности. Подстановочные алгоритмы 234
Использование оценок максимального правдоподобия 234
Использование байесовских оценок 235
Моделирование алгоритмов 237
5.4.2. Распознавание образов в условиях непараметрической неопределенности. Использование оценок плотностей распределения 240
Использование оценок на основе метода Парзена 241
Использование оценок на основе метода k-ближайших соседей 242
Моделирование алгоритмов 243

5.5. Предварительная обработка статистических признаков распознавания 249
5.5.1. Метод главных компонент и отбор информативных признаков 249
Моделирование алгоритма 253
5.5.2. Декоррелирующие свойства дискретных спектральных преобразований....258
Моделирование алгоритма 261
5.5.3. Линейный дискриминантный анализ 264

Глава 6. Распознавание образов в рамках детерминистского подхода 269
6.1. Распознавание образов с использованием функций расстояния 269
6.1.1. Метрические алгоритмы при использовании одного или нескольких эталонных описаний....270
6.1.2. Обучение метрических алгоритмов 275
6.2. Нелинейные преобразования и спрямляющие пространства. Метод потенциальных функций....278
6.2.1. Нелинейные преобразования и спрямляющие пространства. Ядра скалярных произведений....278
6.2.2. Метод потенциальных функций 283
Случай двух классов....283
Случай многих классов 285
Моделирование алгоритма 285
6.3. Метод опорных векторов 291
6.3.1. Случай линейно разделимых классов 291
Случай безошибочно линейно разделимых классов 292
Случай линейной разделимости классов с ошибками....294
Моделирование алгоритма 297
6.3.2. Случай линейно не разделимых классов 299
Моделирование алгоритма 300
6.4. Композиционные методы и алгоритмы распознавания образов 303
6.4.1. Деревья решений и композиции «случайный лес» 306
Деревья решений....306
Показатель загрязненности....307
Расщепление деревьев 308
Усечение деревьев....310
Моделирование алгоритма 310
Случайный лес (Random Forest) на основе баггинга 314
Моделирование алгоритма 316
6.4.2. Композиции базовых алгоритмов с обучением на основе бустинга 319
Моделирование алгоритма 321
6.5. Нейронные сети и их использование для построения алгоритмов анализа данных 323
6.5.1. Типовая архитектура нейронных сетей прямого распространения и их обучение....325
6.5.2. Сходимость нейронных сетей к статистически оптимальным алгоритмам 328
Моделирование алгоритмов 331
6.5.3. Проблема переобучения и практические рекомендации....334

Глава 7. Обучение без учителя и кластерный анализ в рамках статистического и детерминистского подходов.....339
7.1. Статистический подход к задаче классификации без обучения. EM-алгоритм 340
Моделирование алгоритма.....343
7.2. Методы и алгоритмы кластеризации образов в рамках детерминистского подхода.....350
7.2.1. Кластеризация при известном числе классов. Алгоритм K-внутригрупповых средних и алгоритм иерархической кластеризации 352
Алгоритм K-внутригрупповых средних (K-means) и его модификации.....352
Моделирование алгоритма 354
Алгоритмы иерархической кластеризации 356
Моделирование алгоритма 358
7.2.2. Критерии оценки числа классов и сравнительный анализ алгоритмов кластеризации в условиях неизвестного числа классов.....361
Моделирование алгоритмов 362

Список литературы.....371

Приложение. Описание электронного архива.....375

Предметный указатель 377

Введение

Обработка информации, анализ данных, машинное обучение... Эти понятия сегодня определяют большой класс научных методов и компьютерных алгоритмов, используемых практически во всех областях разумной деятельности человека. Особенно это относится к сфере создания и эксплуатации сложных информационных, информационно-измерительных и управляющих систем, где решение задач обработки информации и, в частности, задач анализа данных обеспечивает поддержку принятия эффективных решений.

С точки зрения используемой терминологии следует выделить понятие «обработка информации» как наиболее широкое, определяющее все возможные типы преобразований, систематически выполняемых над данными первичных измерений и наблюдений в интересах автоматизации решения задач в науке, технике и технологии. Анализ данных, по нашему мнению, является более узким понятием и отражает такие преобразования, которые непосредственно направлены на интеллектуальную обработку первичных измерительных данных в интересах извлечения знаний об объектах и формирования информации (вторичных обработанных данных), пригодной для принятия решений. Наконец, машинное обучение — базовый подход, реализуемый в ходе анализа данных, который основан на построении математических моделей объектов в контексте решаемой задачи и синтезировании на этой основе компьютерных алгоритмов принятия решений, обладающих способностью к совершенствованию. Здесь следует сказать о том, что для анализа данных не всегда необходимо прибегать к машинному обучению, т. к. модель объекта, служащая для построения алгоритма принятия решения, может быть изначально задана или постулирована.

В этой книге основное внимание направлено на рассмотрение моделей, методов и алгоритмов решения типовых задач анализа данных, используемых в современных и перспективных информационных системах и технологиях. К этим задачам, в первую очередь, относятся задачи оценивания, регрессии, фильтрации, распознавания и кластеризации образов. При изложении указанных вопросов внимание особо акцентируется на применении различных подходов к реализации алгоритмов, вытекающих из используемой модели анализируемых данных (статистической или детерминистской).

В основу издания положены материалы лекций и лабораторно-практических занятий курсов «Технологии обработки информации», «Распознавание образов», «Нейронные сети и генетические алгоритмы», «Моделирование систем», читаемых автором студентам факультета компьютерных наук Воронежского государственного университета.

Указанные материалы, как и материалы любого другого учебного издания, базируются на изложении и обобщении известных результатов современной теории информационно-измерительных систем, статистической теории решений, методов машинного обучения, представленных в многочисленных цитируемых оригинальных и переводных научных изданиях и статьях.

Кроме того, следует особо выделить ряд цитируемых электронных и печатных учебных изданий, посвященных рассматриваемой тематике, с которыми автор постоянно сверялся в своей работе. Это, в первую очередь, курсы лекций, учебники и учебные пособия по анализу данных и машинному обучению, авторами которых являются К. В. Воронцов, Н. Ю. Золотых, А. А. Барсегян с соавторами, А. Е. Лепский и А. Г. Броневич, Л. М. Местецкий, А. Б. Мерков и др.

В то же время, несмотря на то, что имеется большое количество замечательных монографий и учебников, изданных в нашей стране и посвященных рассматриваемым вопросам, требуется постоянная коррекция и адаптация используемого учебно-методического материала. В этом плане настоящее издание имеет и ряд отличительных особенностей, на которые хотелось бы обратить внимание.

Во-первых, эти отличия отражают, вполне естественно, авторские подбор, систематизацию и изложение известных результатов, объединяющих как классические, так и современные работы по рассматриваемой тематике. Отдельные материалы книги базируются на оригинальных результатах, опубликованных в статьях, монографиях и учебных пособиях автора. К ним относятся: материалы системного плана, представленные в вводной первой главе книги, материалы, отражающие результаты в области синтеза и анализа алгоритмов марковской фильтрации, исследования декоррелирующих свойств дискретных спектральных преобразований, вопросы сходимости нейронных сетей к статистически оптимальным алгоритмам, а также ряд результатов, связанных с решением конкретных примеров и задач.

Во-вторых, большое внимание в книге уделяется компьютерному моделированию практически всех рассматриваемых алгоритмов, которые могут использоваться в прикладных задачах анализа данных. Представленные иллюстративные примеры и компьютерные исследовательские модели алгоритмов выполнены в среде MATLAB (лицензия № 876724). Они основаны на использовании как авторских программных решений, так и стандартных компонентов среды. Общая направленность этих примеров и моделей состоит в том, чтобы дать наглядное представление о реализации современных методов и алгоритмов анализа данных и обеспечить читателя платформой для их содержательного исследования и развития. С этой целью в текст включены фрагменты программного кода, позволяющие не только убедиться в практической реализуемости алгоритмов, но и несущие также содержательную информацию, раскрывающую детали их построения и применения.

Следует отметить, что для понимания и освоения примеров программной реализации алгоритмов, представленных в книге, читатель должен иметь начальные навыки работы в среде MATLAB.

Книга состоит из вводной первой главы и двух частей. *Первая часть* охватывает главы со второй по четвертую и посвящена методам и алгоритмам, используемым при решении задач оценивания, регрессии и фильтрации. *Вторая часть* охватывает главы с пятой по седьмую и посвящена изложению методов и алгоритмов распознавания и кластеризации образов.

В *первой главе* вводятся основные понятия и определения общей теории систем с учетом специфики построения и применения информационно-измерительных систем. Анализируется типовая структура систем извлечения информации и систем обработки информации. Рассматриваются задачи анализа данных в системах обработки информации и базовые подходы для их решения. Вводятся основные понятия и определения, используемые при постановке задач оценивания, регрессии, фильтрации, распознавания и кластеризации.

Цель небольшой *второй главы* состоит в том, чтобы напомнить читателю ряд понятий теории вероятностей, которые будут использоваться далее, а также ввести соответствующие обозначения. Даются понятия случайной величины и случайного вектора, а также их стан-

дартные математические описания. Рассматриваются теоретические и практические вопросы компьютерного моделирования случайных величин и случайных векторов.

В *третьей главе* рассматривается постановка и решение задач оценивания и регрессии в рамках статистического и детерминистского подходов, имеющие общую направленность с точки зрения машинного обучения в интересах построения моделей анализируемых объектов. Описываются разнообразные методы и алгоритмы решения задачи оценки статических параметров и оценки неизвестных плотностей распределения. Излагаются основы регрессионного анализа, включая применение методов регуляризации при построении регрессионных моделей данных. Проводится компьютерное моделирование всех рассматриваемых алгоритмов.

Четвертая глава посвящена вопросам фильтрации — оценивания динамических параметров, описывающих изменяющиеся состояния анализируемых объектов. Рассматривается классическая постановка и решение задач оптимальной нелинейной и линейной фильтрации в дискретном времени, а также нестандартная постановка, возникающая при наличии на входе системы обработки информации пропусков и аномальных наблюдений. Проводится синтез и анализ различных вариантов построения алгоритмов фильтрации. В ходе анализа теоретические оценки ошибок фильтрации сопоставляются с результатами компьютерного моделирования.

В *пятой главе* излагаются основы статистической теории распознавания образов. В рамках байесовской теории принятия решений проводится синтез и анализ множества алгоритмов распознавания, основанных на использовании различных статистических моделей классов образов. Синтез и анализ алгоритмов проводятся при различных уровнях априорной неопределенности, начиная от случая полностью известных статистических описаний классов, вплоть до случая отсутствия таких описаний и использования алгоритмов обучения с учителем. Рассматриваются вопросы предварительной обработки данных и отбора информативных признаков на основе метода главных компонент, использования дискретных спектральных преобразований, линейного дискриминантного анализа.

Шестая глава посвящена вопросам теории и практики построения алгоритмов распознавания в рамках детерминистского подхода при обучении с учителем. Здесь дается описание и исследование разнообразных методов и алгоритмов: в том числе метрических алгоритмов, метода потенциальных функций, метода машин опорных векторов, композиционных алгоритмов на основе технологий баггинга (случайный лес) и бустинга (adaboost и др.), а также нейросетевых алгоритмов обработки информации. С помощью компьютерных моделей на наглядных примерах исследуются, обсуждаются и сопоставляются свойства алгоритмов с точки зрения их эффективности, сложности, способности к обобщению.

Целью *седьмой главы* является краткое изложение основ теории и практики кластерного анализа данных. Дается постановка и решение задачи в рамках статистического подхода (EM-алгоритм) и детерминистского подхода (алгоритмы K-средних и иерархической кластеризации). Особое внимание уделяется задаче кластерного анализа в условиях неизвестного числа классов. Здесь описываются базовые критерии, используемые для оценки числа классов, и проводится сравнительный анализ качества кластеризации, выполняемой на основе различных комбинаций критериев и алгоритмов.

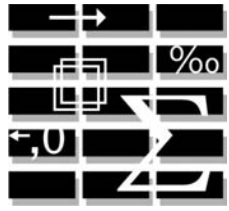
Рассмотренные в книге вопросы, хотя и представляют широкий круг базовых методов решения задач анализа данных, из-за вполне естественного ограничения по объему не включают ряд методов и технологий, используемых в подобных задачах. К ним, например, относятся методы и алгоритмы нечеткой обработки информации, генетические модели и алгоритмы, технологии глубокого обучения (deep learning) в нейронных сетях и других

алгоритмах машинного обучения, специфические методы и алгоритмы, реализуемые для предметной обработки информации (анализ изображений и пр.). Для изучения этих вопросов, при необходимости, можно воспользоваться имеющейся и постоянно обновляющейся литературой.

В заключение следует отметить, что приведенные по тексту и представленные в полном объеме на сайте издательства в виде загрузочных файлов примеры программной реализации алгоритмов обработки информации и их имитационных моделей в среде MATLAB (версии R2012b и выше) могут использоваться в учебном процессе как шаблоны для выполнения лабораторного практикума или служить основой для создания самостоятельных исследовательских проектов при подготовке выпускных квалификационных работ бакалавров и магистерских диссертаций.

Электронный архив

Электронный архив с примерами к этой книге можно скачать с FTP-сервера издательства «БХВ-Петербург» по ссылке **ftp://ftp.bhv.ru/9785977537780.zip** или со страницы книги на сайте **www.bhv.ru** (см. *приложение*).



ГЛАВА 1

Информационные процессы и системы. Обработка информации, анализ данных, машинное обучение

Обработка информации является неотъемлемой составляющей процесса функционирования любой информационной системы. Характер выполняемой обработки может быть различным, но он всегда связан с целевой функцией системы, вопросами обеспечения ее надежности, безопасности, живучести, устойчивости и т. п. Другими словами, процессы обработки информации невозможно изучать в отрыве от изучения методов и моделей общей теории информационных процессов и систем. Поэтому целью предлагаемой вводной главы является рассмотрение базовых определений и понятий этой области, а также особенностей и характеристик информационных систем (и происходящих в них процессов) как особого класса сложных систем, постоянно расширяющих свое присутствие в самых различных сферах нашей жизни. В частности, вводятся понятия процесса обработки информации и технологии обработки информации. Кроме того, на единой методической основе проводится типизация задач, решаемых в системах обработки информации, и, в частности, рассматриваются особенности постановки задач анализа данных и машинного обучения, обеспечивающих выполнение базовых операций в таких системах.

1.1. Основные понятия и определения. Математическое описание систем в рамках теоретико-множественного подхода

Очевидно, что изначально необходимо ввести ряд понятий и определений, которые будут использоваться в ходе изложения материала этой и последующих глав. Прежде всего, предлагается рассмотреть понятие *системы*. Существует достаточно много определений для этого понятия, главная особенность которых состоит в стремлении авторов подчеркнуть некоторые существенные свойства систем (С) как объектов материального мира. По всей видимости, для того чтобы наиболее полно раскрыть понятие С, требуется изначально также дать ряд взаимодополняющих определений для всего спектра сопутствующих понятий и свойств, более подробно рассмотренных в [4, 6, 12, 23, 26, 28, 30, 32, 33, 41, 42, 59].

Система — целостное, органично единое образование, состоящее из множества элементов, находящихся в отношениях или связях друг с другом.

Элемент — простейшая неделимая часть системы, отвечающая предельно детальному рассмотрению системы в рамках решаемой задачи.

Целостность (эмерджентность) — важнейшая характеристика С, которая проявляется в том, что в процессе взаимодействия элементов, входящих в состав системы, появляются принципиально новое качество, свойство, которым не обладает ни один из входящих в С элементов.

Свойство — сторона любого объекта, обуславливающая его отличие или сходство по отношению к другим объектам и проявляющаяся при его взаимодействии с другими объектами.

Качество — свойство или совокупность существенных свойств объекта, определяющих его пригодность для использования по назначению.

Целевое назначение системы (цель системы) — желаемый, потенциально достижимый и фиксируемый результат (исход), который может быть получен в процессе функционирования С.

Эффективность системы — степень соответствия фактического или ожидаемого результата (исхода) процесса функционирования системы желаемому, т. е. степень соответствия цели системы.

Структура — описание совокупности элементов (состав С) и наиболее устойчивых связей (взаимосвязей) между ними (часто без отображения свойств и состояний элементов и связей).

Используются также понятия формальной (функциональной) структуры и материальной (физической) структуры. *Формальная структура* определяет совокупность функциональных элементов (определенных как некоторые функции, преобразующие состояния системы) вместе с их взаимосвязями, необходимую и достаточную для достижения системой поставленных целей. *Материальная структура* определяет конкретную совокупность реальных физических элементов системы и взаимосвязей между ними. При построении системы осуществляется наложение формальной структуры на материальную структуру, которое и определяет так называемую *«амальгированную» структуру* [28], или просто структуру системы. Подобное наложение носит неоднозначный характер, т. к. распределение выполняемых в рамках формальной структуры функций по физическим элементам материальной структуры системы может осуществляться различным образом.

Связь — отношение между элементами С, фиксирующее ограничение их степени свободы, проявляющееся в утрачивании ряда свойств, которыми они потенциально обладают будучи свободными (автономными), и, одновременно, обеспечивающее сохранение структуры и целостности С (т. е. проявление новых свойств, присущих системе в целом).

Взаимодействие систем между собой или их элементов в рамках одной системы — взаимное влияние, приводящее к существенным изменениям в их состояниях (энергетическим, химическим, физическим и т. п.).

Здесь следует отметить, что понятия связи и взаимодействия в определенном смысле идентичные, тем не менее могут отражать различные аспекты отношений между объектами: связь в большей степени направлена на отражение структурных характеристик системы в статике, тогда как понятие «взаимодействие» часто используется для описания изменений свойств, имеющих определенную динамику.

Процесс — целостная совокупность последовательных изменений состояния системы, обеспечивающая реализацию функционального назначения или достижение общей цели системы.

Декомпозиция — разъединение системы на части в интересах ее исследования или проектирования с последовательным самостоятельным рассмотрением этих частей.

Подсистема — относительно независимая часть системы, объединяющая элементы, выделенные при декомпозиции, и реализующая выполнение той или иной функции (подцели, функциональной операции), обеспечивающей достижение общей цели системы.

Названием «подсистема» в этом определении подчеркивается, что выделяемая часть сама обладает основными чертами системы и, прежде всего, целостностью, а также наличием своей цели, являющейся подцелью по отношению к цели исходной С.

Важнейшим понятием для исследования систем является также понятие «внешняя среда».

Внешняя среда — совокупность элементов (объектов) естественного или искусственного происхождения, не входящих в состав системы, но оказывающих на нее определенное воздействие и определяющих существенные условия ее функционирования.

Выделение внешней среды для рассматриваемой системы является весьма важной процедурой, т. к. избыточность или недостаточность в описании внешней среды мешают правильному пониманию сущности С и адекватному анализу ее свойств.

Для теории систем весьма важным является понятие *сложности системы*. Существует ряд подходов к разделению систем на простые и сложные: по количеству входящих в них элементов, по сложности выполняемых функций, по возможности формализованного описания поведения систем и т. п. В частности, академик А. И. Берг определял сложную систему как объект, который можно адекватно описать не менее чем на двух математических языках (например, с помощью аппарата дифференциальных уравнений и аппарата булевой алгебры). Иначе говорят, что такой объект обладает «гибридным поведением». Очень часто сложную систему определяют как систему, которую вообще нельзя корректно описать математически в силу наличия в ней большого количества элементов, связанных неизвестным образом, и/или наличия неопределенностей относительно протекающих в системе процессов и явлений [28]. Для простой системы, напротив, характерна возможность корректного и законченного описания в рамках единого математического аппарата, что позволяет получить аналитические или численные решения, определяющие ее свойства.

В дальнейшем изложении *сложной системой* мы будем называть систему, которой присущи следующие признаки:

- ❒ целевое назначение имеет многоаспектный, многофункциональный, а часто, и слабо формализованный характер;
- ❒ имеется значительное количество разнородных элементов, взаимодействующих друг с другом «непростым» образом;
- ❒ состоит из достаточно самостоятельных подсистем, объединяющих элементы, имеющих свое целевое назначение и решающих свои задачи;
- ❒ присутствуют случайные и не случайные факторы, влияющие на достижение цели системы, воздействие внешней среды трудно предсказуемо, ее элементы в полном объеме выделены быть не могут и остаются неучтенные воздействия;
- ❒ отсутствует единое формализованное описание системы и требуется использование разнородных «языков» формализованного описания, отражающих различные аспекты ее представления и гибридного поведения.

Помимо этого в литературе присутствует термин *большая система*. При этом во многих случаях термины «сложная» и «большая» система используются как синонимы. В ряде источников термин «большая система» относят к системам, имеющим значительное число элементов (чисто количественный аспект).

Мы остановимся на следующем определении и *большой системой* станем называть пространственно-распределенное сообщество взаимосвязанных сложных подсистем, обладаю-

щих определенной степенью автономности, объединенных между собой энергетическими, материальными и информационными связями для обеспечения целенаправленного функционирования как единой системы и использующих человека в процессе организации и управления этими подсистемами и системой в целом.

С понятиями сложной системы и большой системы тесно связано понятие иерархии.

Иерархия — расположение частей и элементов целого в порядке от «высших к низшим». При иерархическом описании сложных систем применяются различные способы [23]. Во всех них используется понятие *уровень иерархии*, имеющее различный смысл в зависимости от способа описания. Существует три основных типа иерархии:

- на основе выделения страт — уровней абстрагирования (аспектов) при описании системы;
- на основе выделения слоев — уровней сложности принимаемых решений при достижении глобальной цели;
- на основе выделения эшелонов или уровней в организационной структуре системы, имеющей четко выделенные семейства управляющих и подчиненных элементов.

Обобщая введенные понятия, можно сделать вывод, что любая С определяется тремя основными категориями: элементы, отношения, свойства. Однозначное и полное задание (описание) этих категорий фактически определяет систему: ее целевое назначение, структуру, качества. Рассмотрение способов формализованного описания С однозначно связано с понятием модели.

Модель — это объект-заменитель объекта-оригинала, обеспечивающий изучение наиболее существенных в интересующем нас аспекте свойств оригинала и, наоборот, позволяющий абстрагироваться от его несущественных в рамках данного рассмотрения свойств. Как правило, объект-заменитель представляет формализованное в рамках некоторого языка (например, математического) описание С. Иногда в этом плане употребляют термин *модель функционирования*, который обозначает модель, обеспечивающую отображение изменения состояний системы во времени, т. е. ориентированную на описание динамики ее функционирования как организованного процесса.

При формализации описания систем (элементов систем) и, вообще, при реализации многих способов моделирования различают два подхода:

- *структурный* — в рамках которого сосредотачиваются на анализе и воспроизведении в модели внутренних свойств и состояний системы;
- *функциональный* — при котором сосредотачиваются на изучении и воспроизведении внешних проявлений системы как преобразователя «вход-выход», а сама система при этом рассматривается как «черный ящик».

Кроме того, для формализованного описания систем часто используется *структурно-функциональный подход*, объединяющий черты предыдущих. Так, при рассмотрении вопросов синтеза и анализа сложных систем обычно проводится их декомпозиция или, как еще часто говорят, структуризация на системы низшего уровня, которые и называют подсистемами и которые, в свою очередь, обладают функциональной целостностью, выполняя определенные операции, обеспечивающие достижение общей цели.

Дальнейшая декомпозиция подсистем приводит к образованию многоуровневой иерархии, которая предполагает не только разделение систем и подсистем на части, но и использование различных аспектов описания системы, понимания ее как объекта материального мира. Подобную декомпозицию, предполагающую последовательное выделение составных частей

системы, реализующих необходимый спектр функций (функциональных операций) как выделенных частей деятельности системы, мы будем называть *структурно-функциональной декомпозицией.*

Напомним некоторые определения, используемые при описании систем и моделей систем в рамках теоретико-множественного подхода.

Декартовым произведением множеств $A \times B$ является множество всех упорядоченных пар (a,b), где $a \in A;\ b \in B.$ Введем понятие отношения $R \subset A \times B$ как подмножества декартова произведения двух множеств, образующих совокупность упорядоченных пар их элементов. Такое отношение, определенное на декартовом произведении двух множеств, называется *бинарным отношением* (БО). Для элементов $a \in A$ и $b \in B$ обозначение aRb следует понимать как $(a,b) \in R$ (a находится в отношении R к b). Важным и часто встречающимся типом бинарных отношений является отношение эквивалентности, которое удовлетворяет следующим трем условиям: aRa — рефлексивность, $aRb \to bRa$ — симметричность, $aRb \cap bRc \to aRc$ — транзитивность.

В общем случае может быть задано декартово произведение n множеств. На нем аналогичным образом задается n-арное отношение как подмножество декартова произведения n множеств. Тогда в соответствии с теоретико-множественным подходом общая система S, заданная на семействе множеств $\overline{X} = \{X_i, i \in I_n\}$, где I_n — множество индексов, определяется как n-арное отношение на этом семействе или подмножество декартова произведения вида:

$$S \subset X_1 \times X_2 \times ... \times X_n . \qquad (1.1)$$

Это общее определение в зависимости от того, какой смысл и какая дополнительная структуризация закладывается в $X_i,\ i \in I_n$, позволяет получить содержательное определение системы в соответствии с тем или иным подходом. При этом элементы $x^{(i)} \in X_i,\ i \in I_n$ сами могут рассматриваться как множества, наделенные определенной структурой.

Так, например, известно, что даже в условиях предельно нечетко выраженного вербального (словесного) описания систему можно представить в виде $S \subset A \times B,$ где A — множество лингвистических переменных (денотат), определяющих объекты рассмотрения, а B — множество функторов (лингвистических переменных, определяющих отношения или формы связи между ними). В этом плане система является отношением на A и B, т. е. подмножеством правильных, наделенных смыслом высказываний, описывающих содержательную взаимосвязь между объектами.

Рассмотрим другой пример. Введем понятие *универсум*, под которым будем понимать базовое множество, содержащее все элементы (объекты) какой-либо области материального или духовного мира. Пусть M — универсум элементов, T — универсум отношений между ними, V — универсум свойств, реализуемых на этих элементах. Тогда система S определяется как

$$S \subset M \times T \times V = Y_S, \qquad (1.2)$$

где Y_S — универсум данного класса систем. Определяя $W = M \times T$ как универсум структурных характеристик или структур, а V — как универсум параметров систем, которые обеспечивают необходимый уровень детализации элементов и связей между ними, без чего невозможно информативное описание системы, получим, что это определение соответствует ранее введенной терминологии. Таким образом, в рамках данного определения системы

определяется подмножество «разумных» (работоспособных) вариантов структуры и параметров, которые могут быть выделены в рамках всех существующих комбинаций.

Введенное ранее общее определение системы (1.1) позволяет также перейти и к реализации функционального подхода. Первым шагом в этом направлении является определение двух множеств: входных объектов X и выходных объектов Y. Тогда система может быть определена как бинарное отношение:

$$S \subset X \times Y. \tag{1.3}$$

При реализации функционального подхода в чистом виде вместо отношения используется отображение, или функция:

$$S : X \to Y. \tag{1.4}$$

Отображение — это такой тип отношений, для которого выполняется:

$$xRy_1 \cap xRy_2 \to y_1 = y_2,$$

или, иначе:

$$(x, y_1) \in R \text{ и } (x, y_2) \in R \to y_1 = y_2.$$

Определенная таким образом система называется *функциональной*. Иногда еще такое отношение обозначается как $f : X \to Y$, причем если xRy, то этот факт обозначается как $y = f(x)$. Отображение $f : X \to Y$ является инъективным, если при $x_1 R y_1$, $x_2 R y_2$ и $y_1 = y_2$ выполняется:

$$y_1 = f(x_1) = y_2 = f(x_2) \to x_1 = x_2.$$

Необходимо отметить, что содержательная основа определений (1.3) и (1.4) отображает, в сущности, различные подходы. В первом случае предполагается установление некоторого правила отбора «разумных» сочетаний (x, y) внутри множества $X \times Y$ при рассмотрении этого множества в целом. Напротив, строго функциональному определению в (1.4) соответствует задание системы, фактически определяющее правило получения для любого $x \in X$ некоторого конкретного $y \in Y$.

Определение системы как n-арного отношения на абстрактных множествах является наиболее общим. В кибернетике, теории управления для составления формализованного описания систем, используются различные математические подходы: графы, дифференциальные уравнения, сети Петри и т. д. Тем не менее, все эти описания в рамках теоретико-множественного подхода могут рассматриваться как отношения вида (1.1)–(1.4).

Дальнейшее углубление и обобщение введенных понятий и характеристик связано с изучением общих закономерностей, объединяющих фундаментальные свойства сложных систем в следующие группы: закономерности взаимодействия части и целого (целостность, аддитивность, интегративность), закономерности иерархической упорядоченности (коммуникативность, иерархичность), закономерности осуществимости (экфинальность, закон необходимого разнообразия, закономерность потенциальной эффективности), закономерности развития (историчность, самоорганизация), закономерности целеобразования и т. д. Их подробный анализ имеется в литературе по теории систем [4, 6, 12, 23, 26, 28, 30, 32, 33, 41], которая может быть при необходимости использована читателем.

1.2. Классификация систем. Информационные системы, информационные процессы, информационные технологии

Цель любой классификации состоит в том, чтобы ограничить выбор подходов к исследуемой системе (объекту), сопоставить выделенному классу объектов адекватные приемы и методы анализа, сформировать правильное отношение к определению эффективности системы. При этом надо помнить, что любая классификация всегда условна, относительна и служит, главным образом, текущим потребностям проводимого исследования, направленного на упорядоченное представление знаний о свойствах изучаемых объектов по выбранным классификационным признакам.

Традиционно [4, 6, 12, 23, 26, 28, 30, 32] рассматривают следующие аспекты классификации систем:

- по происхождению — естественные, искусственные;
- по природе элементов — абстрактные (отображаемые в абстрактных моделях), реально существующие (физические);
- по наличию взаимосвязей с внешней средой — открытые, закрытые (замкнутые);
- по наличию случайных факторов — детерминированные, стохастические;
- по характеру поведения — управляемые, неуправляемые, с целенаправленным поведением, с отсутствием целенаправленного поведения;
- по степени сложности — простые, сложные, большие;
- по степени изменчивости — статические, динамические, развивающиеся;
- по степени организации — хорошо организованные, плохо организованные, самоорганизующиеся;
- по степени участия человека — технические (автоматические), организационно-технические (человеко-машинные, автоматизированные), организационные.

Исходя из представленных общих признаков классификации систем, можно теперь попытаться перейти к определению информационных систем как особого класса систем искусственного происхождения, являющихся одним из предметов рассмотрения. Для этого предварительно введем три основополагающих понятия: *данные*, *информация*, *сигналы*.

Данные — факты, характеризующие объекты, явления или процессы некоторой предметной области и зафиксированные на каком-либо материальном носителе в виде необработанных результатов измерений и наблюдений.

Информация — сведения об объектах, явлениях или процессах в некоторой предметной области, получаемые путем анализа и обработки данных и пригодные для принятия решений потребителем.

Главной особенностью данных является их первичный, исходный характер и, в то же время, их связь с конкретными потребностями пользователей (их востребованность), т. к. они всегда добываются целенаправленно и являются «сырьем» для последующего анализа и обработки.

Согласно приведенному в ГОСТ 15971-90 определению, данные — информация, представленная в виде, пригодном для обработки автоматическими средствами при возможном участии человека.

Таким образом, информацию можно рассматривать как некоторые «вторичные», преобразованные данные, имеющие определенный уровень обобщения и смыслового содержания в интересах решения конкретной задачи в искусственных системах, а данные — как первичную информацию, представленную в виде, пригодном для последующей обработки.

В теории информации применяются различные варианты формализованного определения понятия «информация», которые используют разные количественные меры. Их соотношение и содержательная трактовка подробно рассмотрены в соответствующей литературе [например, 41]. Следует также отметить два различных подхода к пониманию информации: как субъективной категории, возникновение и существование которой связано с конкретными потребностями субъекта, и как некоторой объективно существующей субстанции, например в виде так называемого «информационного поля», существование которой не зависит от ее потребления субъектом.

Приведенные здесь определения придерживаются первого, хотя и более утилитарного подхода, согласно которому данные и информация существуют только по поводу конкретных запросов (интересов) потребителя (пользователя) в связи с необходимостью их использования для принятия конкретных решений при управлении. Нет потребителя — нет и информации [37].

Понятия «данные» и «информация», как также следует из приведенных определений, неразрывно связаны с их физическим носителем. При этом разделяют физические носители, используемые для фиксации и хранения данных и информации, и физические носители, используемые для передачи данных и информации между объектами, или «сигналы».

Сигнал — определенным образом структурированный физический носитель данных и информации при их передаче от одного объекта к другому или от объекта к потребителю.

Физическая природа сигналов может быть различной (электромагнитное излучение, акустическое или механическое воздействие и т. п.). При этом она всегда связана с наилучшим в определенном смысле способом передачи информации в рамках конкретной задачи и ситуации.

Введенные определения позволяют перейти к определению информационной системы как системы, обладающей набором перечисленных ранее классификационных признаков.

Информационная система (ИС) — сложная человеко-машинная система, целевое назначение, элементный состав и структура которой ориентированы на различного рода преобразования данных и информации в интересах обеспечения потребностей пользователей (физических лиц, организаций, органов управления и т. п.).

В состав ИС обычно входят весьма разнородные элементы: средства вычислительной техники, средства добывания данных, средства связи, программные средства, информационные ресурсы, а также обслуживающий персонал.

В ряде источников под информационной системой понимается, в сущности, система документов, их электронных эквивалентов и программное обеспечение, реализующее полный цикл обработки этих документов на компьютере [28]. Мы далее, все же, будем придерживаться данного ранее более общего определения, в котором ИС рассматривается как система, элементами которой являются не только программные средства, но и технические средства, обеспечивающие реализацию ее целевого назначения.

Основным видом взаимодействия в ИС между ее элементами, а также взаимодействия ИС с другими системами и объектами, является информационное взаимодействие. При этом главные функции, связанные с различного рода преобразованиями данных и информации,

выполняют элементы радиоэлектронной аппаратуры (как базовые физические элементы системы).

Информационное взаимодействие, в какой бы форме оно ни осуществлялось, как правило, рассматривается в виде некоторого процесса, т. е. имеет четко выраженную динамику. Это позволяет нам определить понятие информационного процесса.

Информационный процесс — целенаправленно организованный процесс изменения информационных состояний системы, в ходе которого осуществляется четко выраженная последовательность операций, действий по преобразованию информации, в результате которой она может изменять свою форму и/или содержание в пространстве и/или во времени.

Информационные процессы могут быть различных видов [26]: простые (последовательные) и сложные (параллельные, с ветвлением, с обратной связью), однородные и неоднородные (в смысле используемых ресурсов системы и порядка обслуживания), основные и вспомогательные. Но в любом случае информационный процесс всегда есть целенаправленная совокупность операций преобразования данных и информации, реализуемых в определенной физической среде с использованием выделяемых ресурсов системы. Элементарными действиями в каждом процессе являются операции преобразования данных (информации), являющиеся типовыми звеньями в общей последовательности выполняемых изменений информационных состояний.

Базовыми операциями преобразования данных и информации, которые могут использоваться при реализации типовых процессов, являются: измерение, регистрация, сбор, накопление, поиск, фильтрация, сортировка, распределение, анализ, генерация, воспроизведение, отображение и т. п. Для выполнения каждой такой операции используется конкретный алгоритм обработки информации (АОИ), а для реализации информационного процесса в целом — алгоритм функционирования информационной системы (АФИС).

Алгоритм — инструкция, определяющая последовательность выполнения действий при переработке исходного материала в требуемый результат. В нашем случае таким материалом являются данные. Алгоритм всегда определяет представленную в стандартной форме совокупность точных предписаний, понятных человеку и/или компьютеру, детально описывающую последовательность действий, направленных на достижение конкретного результата.

При создании любой ИС необходимо отобразить сущностные основы информационных процессов как совокупностей базовых операций. Воспользуемся для этой цели принятой в [26] формализацией, согласно которой любая обладающая потребительским качеством информационная единица (единица данных) — I характеризуется содержанием — S, формой — F, пространственным размещением — L и временем — T, т. е. набором $I = \{S, F, L, T\}$. Каждая из этих характеристик в ходе выполнения базовой операции и реализации ИП в целом может изменяться. При этом различают следующие виды преобразования данных и информации:

- преобразование содержания, в результате которого получается новая информация (например, получение математической модели объекта или принятие решений на основе регистрируемых первичных данных);
- преобразование формы (например, шифрование, дешифрация информационных сообщений, отображение информации в удобном виде);
- преобразование в пространстве (например, сбор и накопление данных и информации, полученных в разных точках пространства, перенос информации от одного объекта к другому);

- преобразование во времени (например, накопление информации путем объединения данных, полученных в разные моменты времени, хранение данных и информации).

Отражая сущностные информационные основы взаимодействия объектов, систем и элементов систем, указанные преобразования реализуются в действующих системах в различных комбинациях, имея как доминирующее, так и вспомогательное значение, выполняются последовательно и параллельно, образуя достаточно сложный «ансамбль» процессов, подпроцессов и более простых действий (операций). Естественно, что физическая структура системы, обеспечивая протекание информационных процессов, должна оптимизироваться с учетом этих факторов.

Исходя из сказанного можно отметить, что существуют типовые виды информационных процессов, которые группируются по назначению и перечню выполняемых функциональных преобразований данных и информации. Однако конкретная реализация ИП каждого вида может быть различной. Это и позволяет перейти к понятию информационных технологий, которые мы понимаем как способы реализации типовых информационных процессов.

В ряде источников, например [38], под информационной технологией понимают «процесс, использующий совокупность средств и методов сбора, накопления, обработки и передачи данных (первичной информации) для получения информации нового качества о состоянии объекта, процесса или явления (информационного продукта). Этот процесс состоит из четко регламентированной последовательности выполнения операций, действий, этапов разной степени сложности над данными, хранящимися на компьютерах. Основная цель информационной технологии состоит в том, чтобы в результате целенаправленных действий по переработке первичной информации получить необходимую для пользователя информацию».

Таким образом, в этом случае понятия информационного процесса и информационной технологии практически смешиваются. Согласно определению действующего ГОСТ 34.003.90, информационная технология — это приемы, способы и методы применения средств вычислительной техники при выполнении функций сбора, хранения, обработки, передачи и использования данных. Учитывая это, можно предложить далее использовать следующее определение.

Информационная технология — совокупность приемов, способов, мероприятий, обеспечивающих организацию и реализацию информационного процесса конкретного вида с использованием вычислительной (компьютерной) техники, средств сетевого взаимодействия и других технических средств, а также программных средств и информационных ресурсов.

Для того чтобы теперь выделить особый класс информационных систем и технологий — системы и технологии обработки информации, в рамках которых решаются задачи, рассматриваемые в настоящем издании, вернемся к введенной классификации информационных систем и анализу реализуемых в них процессов.

Итак, современные ИС обладают рядом общих признаков сложных и больших организационно-технических систем искусственного происхождения. Основными из этих признаков являются [4, 30, 33] следующие:

- явно выраженная целенаправленность ИС, т. е. наличие совокупности целей (целевых задач), определяющих желаемые результаты, которые должны быть получены в процессе ее функционирования;
- большое количество и разнообразие объектов искусственного и естественного происхождения, с которыми взаимодействуют ИС, — объектов информационного взаимодействия (ОИВ) и, следовательно, разнообразие решаемых ими целевых задач (многофункциональность);

- стохастический характер процессов информационного взаимодействия внутри системы между ее элементами, а также между ИС и объектами внешней среды;
- разветвленность структуры и пространственная распределенность элементов ИС, что определяет сложность реализуемых ИП и необходимость использования сетей информационного обмена данными;
- большие масштабы зоны действия и контура связей ИС с ОИВ, размещаемыми на земле, в воздушном и космическом пространстве, что определяет роль средств телекоммуникации, используемых в составе ИС;
- эволюционный характер процессов создания и модернизации ИС, осуществляемых с непрерывной коррекцией принимаемых технических и технологических решений на основе достижений науки и техники, что позволяет говорить об ИС как о развивающихся системах.

Одновременно ИС характеризуются рядом специфических свойств, которые дают возможность провести дальнейшую классификацию. Основные из этих свойств определяются следующими классификационными признаками: тип объектов информационного взаимодействия, цели и характер информационного взаимодействия с объектами, способы информационного взаимодействия с внешними объектами, общие структурные характеристики и реализуемые в системе процессы информационного взаимодействия. Рассмотрим существо этих признаков подробнее.

1) Объекты информационного взаимодействия могут быть простыми и сложными, естественного и искусственного происхождения и т. д. Каждый из них, в конечном счете, может рассматриваться как система, обладающая конкретными свойствами. Будем также разделять объекты информационного взаимодействия на ОИВ — источники информации и ОИВ — потребители информации, получаемой в результате информационного взаимодействия.

Соответственно ИС можно классифицировать на системы, взаимодействующие с простыми и сложными объектами, естественного и искусственного происхождения, а также на системы, ориентированные на получение информации от объектов внешней среды, и на системы, ориентированные на формирование и предоставление информации объектам внешней среды. Возможна также комбинация, когда ИС взаимодействует с одними ОИВ как с источниками, а с другими ОИВ как с потребителями информации. В качестве примера ИС, получающих информацию от объектов естественного происхождения, можно привести системы мониторинга окружающей среды.

Очевидно, что в чистом виде трудно говорить о наличии «одностороннего» характера взаимодействия с ОИВ (только на получение или только на выдачу информации). Такое деление скорее следует понимать в смысле определения главных целей ИС и соотношения объемов данных и информации на входе и на выходе системы.

2) Информационные системы, как уже отмечалось, относятся к категории целенаправленных сложных систем. Это означает, что они функционируют в соответствии с глобальной целью, задаваемой некой надсистемой высшего уровня. В качестве такой надсистемы может выступать государство, ведомство, организация, общество и сообщество людей. Такая надсистема определяет для ИС перечень целевых или внешнеобусловленных задач, формулируемых на вербальной основе и содержащих ключевые термины, обозначающие объекты и существо информационного взаимодействия. Обычно вербальное определение внешнеобусловленных задач дополняют количественными показателями, задающими те или иные меры качества, при этом вводятся внешнесистемные требования в виде границ допустимых значений для используемых показателей.

Наиболее общая классификация ИС по целевому назначению [4, 6, 12, 18, 23, 26, 28, 30, 32, 33, 41, 42, 59] предполагает выделение следующих основных классов систем:

- системы передачи информации;
- системы извлечения информации;
- системы разрушения информации;
- системы защиты информации;
- системы информационной поддержки управления;
- интегрированные (комбинированные) системы.

Для систем последнего класса характерно, что в их состав (как подсистемы) входят не менее двух специализированных ИС перечисленных здесь классов.

Введенная классификация ИС по целевому назначению тесно связана с определением характера доминирующего для системы информационного взаимодействия (информационного процесса). Кроме того, влияние внешней среды, содержащей большое количество разного рода объектов, включая и не относящиеся к сфере интересов ИС, приводит к возникновению помех и дестабилизирующих факторов, нарушающих информационные взаимодействия. Необходимость борьбы с помехами различного происхождения, а также минимизации влияния дестабилизирующих факторов, определяет необходимость реализации специальных мер защиты и оказывает существенное влияние на структуру ИС. Поэтому облик ИС определяется характером внешнего информационного взаимодействия системы с ОИВ, а также информационного взаимодействия элементов внутри системы. Можно выделить три вида такого взаимодействия: *согласованное*, *индифферентное*, *конфликтное*.

- **Согласованное информационное взаимодействие** подразумевает единство целей, возникающих при функционировании ИС (элементов ИС) и ОИВ. Оно характеризуется наличием достаточно полных априорных сведений об условиях, параметрах и характеристиках реализуемых физических каналов передачи информации, распространяемых в них сигналах и способах кодирования передаваемой полезной информации. Такая ситуация характерна, прежде всего, для систем передачи информации.
- **Индифферентное информационное взаимодействие** реализуется в ситуации «безразличия» участвующих в нем объектов по отношению к процессу получения информации об их состояниях в ИС. При этом уровень априорной неопределенности по сравнению с предыдущим случаем возрастает. Индифферентность наиболее типична для систем мониторинга окружающей среды, некоторых систем дистанционного зондирования, радиолокации и других, относящихся к классу ИС извлечения информации.
- **Конфликтное информационное взаимодействие** отличается наличием антагонизма целей сторон-участников взаимодействия. Конфликтный характер приводит к еще большему уровню априорной неопределенности относительно параметров и характеристик реализуемых каналов передачи информации, а также используемых в них сигналов. Это, как правило, связано с реализацией одной из сторон (обеими сторонами) специальных мероприятий по скрытию или искажению информации, призванных затруднить работу другой стороны. Такая ситуация характерна, например, для функционирования систем радиопротиводействия и радиоэлектронной разведки, относящихся, соответственно, к классам ИС разрушения и извлечения информации [4, 18, 33].

3) Информационное взаимодействие с внешними объектами может осуществляться различными способами, которые можно разделить на два больших класса: *бесконтактные и контактные.*

- **Бесконтактные способы** информационного взаимодействия реализуются, как правило, на основе электромагнитных и других физических полей, выполняющих функцию переносчика информации и распространяющихся через разъединяющую объекты физическую среду. При этом осуществляется модуляция полезными сообщениями некоторых параметров сигналов, излучаемых и распространяемых одной из сторон в спектре электромагнитных волн. Полезные сообщения могут зарождаться как внутри самой ИС, так и в объектах информационного взаимодействия. В первом случае сообщения вносятся в излучение искусственным образом с помощью специальных устройств модуляции параметров сигналов. При таком способе взаимодействия, как правило, реализуется информационный обмен между его участниками. В другом распространенном варианте полезная информация получается непосредственно в процессе воздействия создаваемых в ИС зондирующих электромагнитных волн на объекты (эффекты отражения, поглощения, переизлучения) или при формировании объектами, функционирующими в режиме излучения, собственных электромагнитных волн. При этом происходит естественная модуляция параметров излучаемых сигналов, что позволяет в ИС получить интересующую информацию о состоянии объектов. Подобная ситуация наиболее характерна для систем извлечения информации, действующих по отношению к ОИВ естественного и искусственного происхождения.
- **Контактные способы** предполагают, что элементы системы и объекты информационного взаимодействия в пространстве не разъединены, причем выход системы непосредственно «подключен» ко входу объекта или выход объекта «подключен» ко входу системы. При этом имеется специальный элемент, осуществляющий переход от системы к объекту или наоборот.

4) Общие цели, определяющие назначение ИС, будут достигнуты, и внешнеобусловленные задачи станут успешно решаться, если средства их реализации в виде совокупности входящих в ИС подсистем для обеспечения информационного взаимодействия выполняют определенные функциональные преобразования (внутриобусловленные задачи). Перечень внутриобусловленных задач должен быть достаточным для покрытия каждой из внешнеобусловленных задач и совокупной поддержки функционирования системы в целом. По отношению к каждой внутриобусловленной задаче определяется перечень внутрисистемных частных показателей качества и других характеристик подсистем, для которых определяются соответствующие требования. Это означает, что вводятся внутрисистемные характеристики, объединяющие внутриобусловленные задачи и их количественные описания, которые, в свою очередь, выступают в качестве внешнесистемных по отношению к подсистемам, входящим в состав ИС. Последовательно пройдя подобную декомпозицию задач по отношению ко всем выделенным уровням иерархии рассматриваемой системы, можно получить так называемое *обобщенное дерево функций системы*.

Соответственно, для ИС всегда можно выделить два контура информационного взаимодействия [30].

Первый из них относится непосредственно к реализации целей глобального информационного взаимодействия с объектами внешней среды (внешнеобусловленных задач). Будем называть его *внешним контуром информационного взаимодействия.*

Обычно выделяют следующие целевые задачи внешнего контура информационного взаимодействия:

- обеспечение информационного обмена ОИВ;
- извлечение информации о состоянии ОИВ;

- управление функционированием ОИВ;
- нарушение функционирования ОИВ;
- обеспечение информационной безопасности ОИВ;
- навигационно-временное обеспечение ОИВ.

Как видно, этот перечень внешнеобусловленных задач достаточно хорошо согласуется с ранее введенной классификацией систем по их целевому назначению и видов информационного взаимодействия.

Второй контур называется *внутренним контуром информационного взаимодействия.* Он относится к совокупности информационных процессов и операций, обеспечивающих нормальное функционирование, взаимодействие и координацию работы пространственно-распределенных подсистем (элементов) ИС. Очевидно, что если для внешнего контура характер информационного взаимодействия может быть самым различным (согласованное, индифферентное, конфликтное), то для внутреннего контура информационное взаимодействие, как правило, носит согласованный характер.

В основе функционирования внешнего и внутреннего контуров всегда лежат те или иные информационные процессы (подпроцессы), каждый из которых состоит из определенного набора операций преобразования данных и информации. К таким ИП обычно относят процессы передачи, извлечения, накопления и анализа данных, хранения и обобщения данных и информации и др. Указанные ИП во многом определяют структуру ИС, т. е. состав необходимых подсистем и функциональных модулей, реализующих как операции информационного взаимодействия с ОИВ, так и операции обеспечения функционирования внешнего и внутреннего контуров.

Информационные процессы могут быть *доминирующими*, представляющими основу общего процесса функционирования ИС, и *вспомогательными подпроцессами*, вложенными в какой-либо доминирующий процесс. Но в любом случае они должны быть целостными, отвечающими четко определенному конечному результату их выполнения. Очевидно также, что не все типовые информационные процессы, которые относятся к внешнему контуру, могут быть реализованы во внутреннем контуре. И наоборот, имеется ряд специфических процессов (например, процесс накопления информации), которые имеют сугубо внутренний характер [4].

Таким образом, можно утверждать, что структурные характеристики ИС объективно связаны с необходимым перечнем информационных процессов, разыгрываемых в системе при ее взаимодействии с ОИВ, и их реализацией на основе физических элементов системы (подсистем), выделенных на данном уровне иерархического описания. Далее можно перейти к формированию типового облика конкретной реализации системы, относящейся к одному из выделенных ранее классов. Обычно для этого проводят анализ [26, 28], в ходе которого необходимо определить объекты и формы представления информации, методы и средства ее передачи, обработки и преобразования, исходя из известных внешнеобусловленных и внутриобусловленных задач системы.

Основываясь на выполненной ранее детализации, определим конкретно класс систем извлечения информации.

Система извлечения информации (СИИ) представляет собой сложную целенаправленную, во многих случаях пространственно-распределенную систему (ПРС), предназначенную для бесконтактного и/или контактного извлечения информации о состоянии объектов естественного или искусственного происхождения при реализации согласованного, индифферентного и конфликтного информационного взаимодействия.

В качестве прототипа СИИ можно, например, рассматривать системы радиомониторинга, подробно описанные в [18]. Основными целями функционирования таких систем являются обнаружение и распознавание типов радиоизлучающих объектов внешней среды (ОИВ-источников), оценка частотно-временных параметров излучаемых ими радиосигналов, определение координат объектов на основе объединения информации, добываемой в пространственно-разнесенных пунктах (датчиках) системы.

С этих позиций внешний контур информационного взаимодействия в системах радиомониторинга всегда реализует процесс извлечения информации о состоянии объектов информационного взаимодействия, который обеспечивает результирующие преобразования L, S. В рамках этого процесса осуществляется согласованное, индифферентное или конфликтное информационное взаимодействие, основной целью которого является получение информации о состоянии ОИВ-источника путем принятия соответствующих решений (обнаружение объекта, распознавание объекта, оценивание параметров объекта и др.). Эта информация получается в системе на основе измерения, регистрации, накопления и анализа данных (первичной информации) о параметрах принимаемых сигналов (при бесконтактном способе взаимодействия). Сигналы могут быть естественного происхождения, т. е. формироваться самим источником пассивно, под влиянием внешних природных факторов (например, облучения независимым внешним источником), или внутренних энергетических возможностей. Эти сигналы также могут специально создаваться на самом объекте информационного взаимодействия искусственным образом, что, в сущности, дает возможность получения необходимой первичной информации для СОИ.

Представленная на рис. 1.1 схема последовательности базовых операций и внутренних подпроцессов показывает, что доминирующий здесь процесс извлечения информации имеет ветвление и обратную связь. В данном случае это отражает два возможных режима его реализации: режим «обучения» системы, т. е. режим изучения объекта и накопления информации, характеризующей его поведение, которая в дальнейшем может использоваться при принятии решений относительно состояний объекта, интересующих потребителя, а также второй режим, который реализует непосредственное принятие решений с учетом ранее полученной информации. Обратная связь характеризует возможность совмещения указанных режимов реализации процесса, а также — управления процессом извлечения информации по результатам принимаемых решений. Классическим примером систем, в которых реализуется такой ИП, являются системы пассивной радиолокации, системы радиоразведки, системы мониторинга окружающей среды.

В рамках доминирующего ИП, объединяя базовые операции, действуют основные вложенные подпроцессы: подпроцесс приема, измерения и регистрации параметров сигналов — подпроцесс первичной обработки сигналов (результирующее преобразование F), подпроцесс передачи информации, обеспечивающий обмен данными и информацией между различными элементами системы (результирующее преобразование L), подпроцесс сбора, накопления и анализа данных — подпроцесс вторичной обработки измерительной информации, обеспечивающий формирование информации для принятий решений (результирующие преобразования F, T, S), подпроцесс хранения данных и информации, обеспечивающий формирование хранилищ данных и информации и их предоставление пользователю в виде, пригодном для поддержки принятия решений (результирующее преобразование T).

В этой схеме уместно особо выделить подпроцесс сбора, накопления и анализа данных, или просто анализа данных. Этот подпроцесс во многих случаях имеет самостоятельное значение (например, при контактном информационном взаимодействии) и является доминирую-

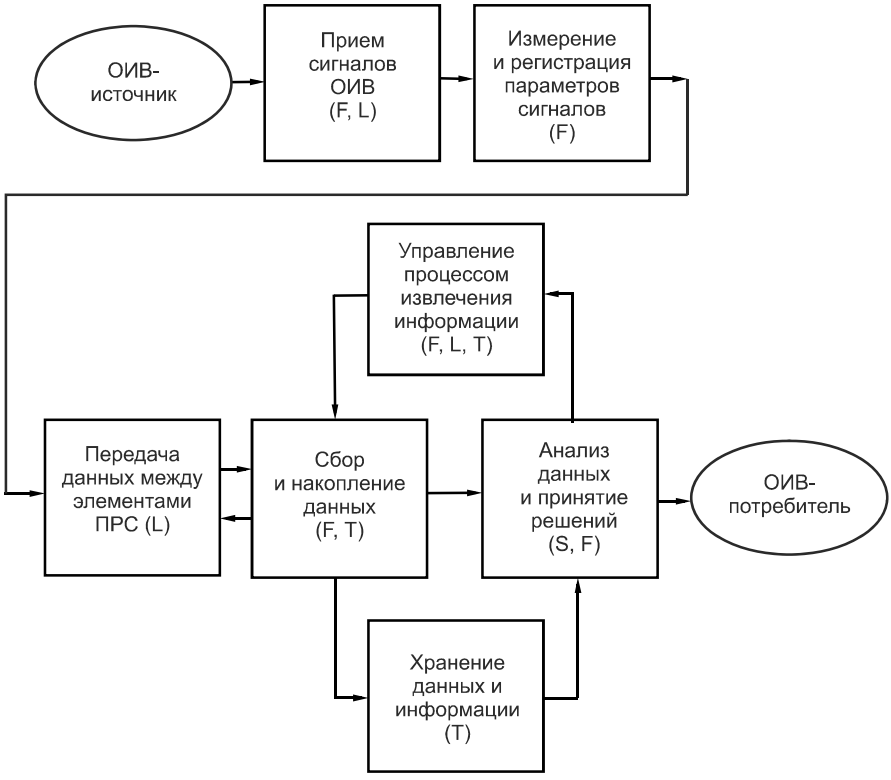


Рис. 1.1. Типовая схема процесса извлечения информации в системах радиомониторинга

щим для отдельного класса систем, которые называют *системами обработки информации* (СОИ). В связи с этим следует привести следующие определения.

Обработка информации (ОИ) — систематическое выполнение операций над данными, представляющими предназначенную для обработки информацию (ГОСТ 15971-90).

Система обработки информации (СОИ) — совокупность технических средств и программного обеспечения, а также методов обработки информации и действий персонала, обеспечивающая выполнение автоматизированной обработки информации (ГОСТ 15971-90).

Типовая схема организации процесса обработки информации, включающего характерные для него подпроцессы и операции преобразования данных и информации, представлена на рис. 1.2.

Целевое назначение этого процесса — проведение содержательного анализа и обобщения большого объема данных, получаемых от ОИВ-источников, на основе решения либо вычислительной задачи, либо задачи принятия информационных решений. В результате решения задачи формируется новая информация относительно прошлого, текущего или будущего состояния объектов, которая интересует потребителя.

Главное отличие этой схемы от предыдущей состоит в том, что здесь осуществляется регистрация (ввод) данных в систему в режиме согласованного и контактного взаимодействия (например, путем непосредственного ввода данных в систему), а условия решения вычислительной задачи или задачи принятия решений заранее известны и более или менее определены.

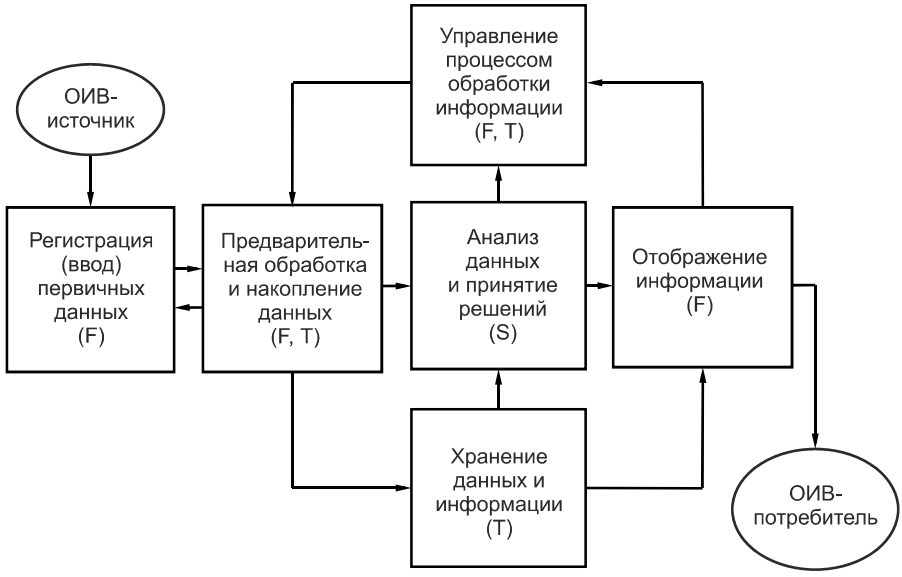


Рис. 1.2. Типовая схема процесса обработки информации

Таким образом, мы видим, что процесс обработки информации может быть или вложенным в процесс извлечения информации, или главным, доминирующим ИП, а реализующая его система обработки информации может являться подсистемой более сложной системы или быть самостоятельной. Как доминирующий, процесс ОИ реализуется, например, в статистических информационных системах, системах поддержки принятия решений (СППР) и т. п. [7, 41], ориентированных на решение структурированных задач, для которых имеются исходные данные и известны алгоритмы, ведущие к их решению.

Так, в [7] определена типовая структура СППР, которые определяются как системы обработки информации, предназначенные для анализа данных, относящихся к определенной предметной области с целью поиска решений. В этой структуре выделены три подсистемы: подсистема ввода данных, подсистема хранения данных и подсистема анализа данных, являющаяся главной. При этом в зависимости от степени сложности решаемой задачи анализа данных (степени «интеллектуальности» анализа) выделены следующие разновидности СППР: системы информационно-поискового анализа, подсистемы оперативного анализа, подсистемы интеллектуального анализа данных.

Определив таким образом процесс обработки информации, можно, по аналогии, дать следующее определение реализующей его информационной технологии.

Технология обработки информации — совокупность приемов, способов, мероприятий, обеспечивающих организацию и реализацию процессов извлечения информации о состоянии объектов при бесконтактном или контактном взаимодействии с использованием вычислительной (компьютерной) техники, средств сетевого взаимодействия и других технических средств, а также программных средств и информационных ресурсов.

1.3. Задачи анализа данных в системах обработки информации и базовые подходы для их решения

Мы видим, что процесс обработки информации включает комплекс разнообразных подпроцессов и операций. Сосредоточимся в дальнейшем на рассмотрении вопросов анализа данных, как базового подпроцесса, реализуемого в СОИ.

Анализ данных в широком смысле [38] — область науки, занимающаяся построением и исследованием математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (наблюдаемых) данных, процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.

Анализ данных имеет несколько аспектов, основан на применении множества подходов и охватывает разнообразные методы в различных областях науки и деятельности. Отражением этого является, в частности, то, что в многочисленных источниках [7, 10, 16, 17, 22, 32, 35, 52, 55] используются ряд близких и сопутствующих понятий. К ним относятся такие понятия, как *искусственный интеллект*, *машинное обучение*, *Data Mining.*

Термин **искусственный интеллект** (artificial intelligence, AI) трактуется как научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования видов человеческой деятельности, традиционно считающихся интеллектуальными. Соответственно, искусственный интеллект определяется как способность автоматических систем (не обязательно компьютерных) брать на себя отдельные функции интеллекта человека. Таким образом, искусственным интеллектом (ИИ) называют свойство систем выполнять творческие функции, которые традиционно считаются прерогативой человека.

Современный ИИ делится на множество различных направлений, из которых два основных: прикладной (applied AI) и сильный ИИ (strong AI). *Прикладной ИИ* изучает возможности использования компьютеров и программного обеспечения для решения конкретных задач, которые не выходят за рамки познавательных способностей человека. Большинство современных систем ИИ относится к этой категории. Синонимы: узкий (ограниченный) ИИ (narrow AI) и слабый ИИ (weak AI). К системам прикладного искусственного интеллекта относятся системы компьютерного зрения, системы распознавания речи и рукописного текста, многочисленные робототехнические системы и т. п. Все эти задачи, по сути, относятся к задачам интеллектуального анализа данных.

Следует отметить, что *сильный искусственный интеллект* — перспективное направление в развитии ИИ, ставящее своей целью создание систем, сравнимых по своим возможностям с интеллектом человека или превосходящих его. Компьютер с сильным ИИ должен быть в состоянии решать любую интеллектуальную задачу, которую только способен решить человек.

Машинное обучение (machine learning) в известных источниках характеризуется как методология получения новых знаний на основе алгоритмов и программ, реализованных с использованием современных средств вычислительной техники. Единообразного определения машинного обучения на сегодняшний день нет. Например, можно сослаться на следующие трактовки. *Машинное обучение* — процесс, в результате которого машина (компьютер) способна показывать поведение, которое в нее не было *явно* заложено (запрограммировано) (*A. L. Samuel*, Some Studies in Machine Learning Using the Game of Checkers // IBM Journal.

July 1959. P. 210–229). Т. М. Митчелл дал такое определение: «Машинное обучение — это наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время работы» или «Говорят, что компьютерная программа *обучается* на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта E» (*T. M. Mitchell*, Machine Learning. McGraw-Hill, 1997).

В [10] дается следующая трактовка этой области знаний: теория обучения машин (machine learning, машинное обучение) находится на стыке прикладной статистики, численных методов оптимизации, дискретного анализа и за последние 50 лет оформилась в самостоятельную математическую дисциплину.

Методы машинного обучения составляют основу еще более молодой дисциплины **Data mining** (в [7], например, утверждается, что машинное обучение обозначает все технологии Data Mining). Термин Data Mining получил свое название на основе двух понятий: поиска ценной информации в хранилище (базе) данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей. Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, обнаружение знаний в базах данных и т. п. [7, 55].

В соответствии с известным определением [7] Data Mining — исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых первичных данных скрытых знаний (в нашем понимании — содержательной информации), которые ранее не были известны, практически полезны, доступны для интерпретации человеком. Особо подчеркивается, что:

- ❒ знания должны быть новые, ранее неизвестные;
- ❒ знания должны быть нетривиальны (содержать неочевидные выводы);
- ❒ знания должны быть практически полезны;
- ❒ знания должны быть доступны для понимания человеком.

Очевидна близость всех рассмотренных понятий, из которых понятия «анализ данных» и Data Mining, являясь практически синонимами, содержат как теоретический, так и прикладной аспект, тогда как термины «искусственный интеллект» и «машинное обучение», на наш взгляд, в большей степени определяют методологию решения задач обработки информации в части анализа данных.

Рассмотрим типовые задачи анализа данных в СОИ, решение которых обеспечивает необходимые преобразования (задачи Data Mining) как базовые операции ИП, являющихся звеньями в общей последовательности выполняемых изменений информационных состояний систем. К числу основных задач в соответствии с различными источниками [2, 3, 7, 9–11, 13, 14, 16, 17, 20–22, 24, 29, 32, 35, 48–52, 55] предлагается отнести:

- ❒ задачу распознавания (классификации);
- ❒ задачу кластеризации;
- ❒ задачу оценивания и регрессии;
- ❒ задачу установления ассоциаций.

Рассмотрим кратко содержание этих задач с учетом терминологии, используемой в различных источниках. Для этого первоначально введем ряд терминов, которые помогут сформировать более точные представления об этих задачах и связях между ними. К ним относятся такие термины, как *наблюдения*, *признаки*, *образы*, *классы образов* и т. п.

Наблюдения — первичные данные, отражающие характеристики объектов, получаемые в процессе измерения и регистрации параметров сигналов (любых носителей информации) на основе используемых в СОИ физических датчиков в ходе бесконтактного или контактного информационного взаимодействия. Наблюдения, как правило, организованы в виде вектора $z = (z_1,...,z_q)^T \in R^q$. Далее, иногда, также будут использоваться термины «наблюдаемые параметры», «измерительные данные».

Признаки (дескрипторы) — особым образом отобранные первичные данные (наблюдения) или данные, полученные в результате их преобразования, которые используются для составления информативного описания объекта с точки зрения решения конкретной задачи обработки информации. Признаки, как правило, организованы в виде вектора признаков (дескрипторов) $x = (x_1,...,x_n)^T \in \mathbf{R}^n$. Обычно размерность пространства признаков существенно меньше, чем размерность пространства наблюдений ($n \le q$), хотя не исключается ситуация, когда наблюдения и признаки совпадают $z \equiv x$. Задача отбора информативных признаков является отдельной задачей, которая может решаться на различных этапах процесса обработки информации.

Состояния — совокупность непосредственно не наблюдаемых параметров объекта, характеризующих интересующие нас существенные свойства объекта, присущие ему в текущий момент времени и, возможно, изменяющиеся в другие моменты времени. В процессе информационного взаимодействия ненаблюдаемые параметры объекта могут быть представлены в виде вектора $s = (s_1,...,s_m)^T \in \mathbf{R}^m$. Состояния объекта часто рассматриваются как некоторое сообщение, которое требуется извлечь в процессе обработки наблюдений, тем или иным образом связанных с состояниями.

Образ — формализованное описание конкретного объекта в пространстве используемых признаков $x = (x_1,...,x_n)^T \in \mathbf{R}^n$.

Класс образов — категория, определяющая совокупность объектов, имеющих определенное сходство, общие свойства, проявляющиеся при их описании в виде образов, и, соответственно, отличающихся по этим свойствам от объектов, включаемых в другие классы.

Множество классов будем далее обозначать $\Omega = \{\omega_1,...,\omega_M\}$. Обозначение $x \in \omega_i$ свидетельствует о том, что этот образ принадлежит i-му классу.

Эталонное описание класса — априорные сведения и характеристики класса в пространстве используемых признаков, опирающиеся на используемую для анализа математическую модель объектов этого класса.

Обучение — процедура получения эталонных описаний классов в рамках системы выбранных признаков, основанная на использовании совокупности обучающих (опытных) данных и априорных сведений относительно физической природы анализируемых объектов. В широком смысле обучение можно охарактеризовать как получение математической модели объектов, описывающей закономерности в данных и используемой в процессе настройки алгоритма анализа в контексте решаемой задачи.

Принятие решения — процедура отнесения объекта, представленного своим образом, к заданному классу или процедура определения неизвестных параметров объекта (состояний), представленного наблюдениями, основанная на использовании ранее полученной математической модели объекта или класса объектов в контексте решаемой задачи обработки информации. В основе принятия решений всегда лежит решающее правило, или алгоритм принятия решений, определяющий конкретную последовательность преобразования входных данных в итоговый результат.

Обучающая выборка — совокупность реализаций (прецедентов) вектора признаков, описывающая конкретное множество объектов в системе этих признаков и используемая при обучении $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(i)} = (x_1^{(i)},...,x_n^{(i)})^T \in \mathbf{R}^n$.

Выделяют ситуацию, когда обучающие данные изначально проиндексированы (помечены), т. е. смешанной совокупности X^N дополнительно придается совокупность индексов $D^N = \{d^{(1)},...,d^{(N)}\}$, $d^{(i)} \in D$. Здесь множество D — конечное множество целых чисел $D = \{1,2,...,M\}$, определяющих принадлежность объектов классам. Иногда здесь для каждого класса образов обучающие данные $X^{N_j} = \{x^{(j,1)},...,x^{(j,N_j)}\}$, $x^{(j,i)} = (x_1^{(j,i)},...,x_p^{(j,i)})^T \in \mathbf{R}^n$, $j = \overline{1,M}$ задаются отдельно.

Кроме того, возможно задание обучающей выборки, в которой X^N придаются необходимые «ответы» $Y^N = \{y^{(1)},...,y^{(N)}\}$ $y^{(i)} \in \mathbf{R}^h$, $i = \overline{1,N}$, каждый из которых является вещественнозначным вектором, т. е. принадлежит конечномерному вещественному векторному пространству (например, евклидову).

Во многих задачах при обучении D^N или Y^N изначально не заданы и имеется только совокупность данных X^N без указания принадлежности образов конкретным классам объектов или определения необходимого ответа.

Тестирующая выборка — совокупность реализаций вектора признаков, описывающая конкретное множество объектов в системе признаков и используемая для проведения тестирования (контрольного эксперимента) и оценки качества ранее полученного (синтезированного) алгоритма обработки информации $X^P = \{x^{(1)},...,x^{(P)}\}$, $x^{(i)} = (x_1^{(i)},...,x_n^{(i)})^T \in \mathbf{R}^n$.

В этом случае данные обязательно должны быть либо проиндексированы (помечены) с помощью совокупности индексов D^N, определяющих реальную принадлежность каждого объекта одному из возможных классов, либо снабжены контрольными ответами Y^N. В противном случае качество используемого алгоритма невозможно оценить.

Опираясь на введенную терминологию, можно теперь перейти к более детальному определению содержания ранее перечисленных типовых задач обработки информации.

Задача распознавания (классификации, различения, узнавания) состоит в отнесении некоторого объекта, описываемого совокупностью характеристик, или признаков, определяющих «образ» этого объекта, к одному из ранее выделенных классов. При постановке этой задачи считается, что количество классов и их эталонные описания известны или могут быть определены в процессе обучения.

Во многих случаях эталонное описание классов может быть получено на основе постулирования некоторой математической модели образов. Например, в задачах распознавания радиосигналов часто используются вероятностные модели, основанные на использовании гауссовского распределения флуктуаций (шумов), описывающих вариации анализируемых образов в пределах каждого класса.

В других случаях, перед тем, как система сможет выполнять ту или иную функцию, предполагается ее обучение на множестве прецедентов обучающей выборки объектов распознавания, в результате чего формируются необходимые эталонные описания. В работе [24] вносится следующее уточнение: задачу распознавания на основе имеющегося множества прецедентов называют *классификацией с обучением (с учителем)*.

Задача кластеризации состоит в разбиении некоторого множества объектов, представленных своими образами, на классы (кластеры, группы), в каждый из которых помещаются

в известном смысле (по степени сходства) «близкие» образы, в то время как образы, помещаемые в различные классы, имеют существенные «отличия». При этом следуют выделить случай, когда число классов известно, и случай, когда число классов неизвестно. Эта задача еще называется *автоматической классификацией* или *классификацией без обучения (без учителя)*.

Задача классификации при не заданном количестве классов является наиболее сложной для решения — высокое качество разбиения образов на классы достигается только при достаточно большей степени различия образов, объективно относящихся к разным классам.

Замечание 1. Понятия наблюдения, признаки, состояния, образы, классы и т. п. всегда определяются в контексте решаемой задачи. Одно и то же может рассматриваться и как образы, и как классы образов. При автоматическом анализе текста часто требуется определять принадлежность каждого символа или группы символов (слова) одному из классов. Например, все печатные буквы русского языка образуют классы. Каждый класс может быть представлен различными графическими изображениями одной и той же буквы. В нашем случае это образы, отличающиеся формой и размерами изображения. Признаками здесь являются параметры формы и размера изображения этих букв. Задача распознавания состоит в отнесении каждого изображения к конкретной букве, независимо от того, в каком формате она набрана. С другой стороны, можно рассматривать в качестве классов алфавиты различных языков, на которых написан текст. Образы теперь — это различные буквы каждого алфавита. Признаками, как и ранее, могут служить параметры графических изображений букв. Задача распознавания в таком контексте состоит в автоматическом определении языка, на котором написан текст.

Задача оценивания состоит в определении неизвестных характеристик объекта на основе анализа (обработки) совокупности первичных данных, представленных либо непосредственно как наблюдения, либо в виде признаков (отобранных наблюдений). В качестве оцениваемых характеристик могут выступать самые различные величины. К ним, например, может относиться один из параметров состояния объекта, либо вектор параметров, характеризующих полное состояние объекта $s \in \mathbf{R}^m$.

Во многих случаях задача оценивания направлена на определение неизвестных параметров статистического распределения (плотности распределения вероятностей, функции распределения вероятностей) известного аналитического представления или на определение вида закона распределения в целом, если его аналитическая форма неизвестна. Такие задачи называются, соответственно, *задачей параметрического оценивания* и *задачей непараметрического оценивания* статистического распределения.

В классе задач оценивания выделяют задачу **фильтрации**, которая состоит в оценивании непосредственно ненаблюдаемой реализации векторного случайного процесса $s(t) = (s_1,...,s_m)^T \in \mathbf{R}^m$, $t \in T$ (случайной функции времени) или реализации случайного поля $s(u,v) = (s_1,...,s_m)^T \in \mathbf{R}^m$, $(u,v) \in \Omega_{uv}$ (случайной функции двух и более переменных), описывающих изменения состояния объекта во времени и/или пространстве.

Для проведения фильтрации используются полученные наблюдения $z \in R^q$, зависящие в данном случае от определяемых состояний объекта $z(t) = z(s(t),t)$, $t \in T$ или $z(u,v) = z(s(u,v),u,v)$, $(u,v) \in \Omega_{uv}$ и индексирующих переменных. Обычно такая зависимость характеризует влияние помех (шумов), определяющих условия проведения наблюдений. Термин *фильтрация состояний объекта*, подчеркивает динамический характер оцениваемого при этом информационного сообщения. Классическая задача оценивания является частным случаем задачи фильтрации, в которой вектор параметров $s(t) = \text{const}$, $t \in T$, т. е. не изменяется в пределах индексирующего множества моментов времени.

Задача регрессии состоит в установлении функциональной зависимости между ожидаемыми значениями зависимых (выходных) переменных и значениями независимых (входных) переменных. Решение задачи базируется на использовании особым образом организованной обучающей выборки данных, в которой задаются совокупность значений вектора независимых переменных $X^N = \{x^{(1)},...,x^{(N)}\}$ и соответствующая ей совокупность значений зависимой переменной — «ответов» обучающей выборки $Y^N = \{y^{(1)},...,y^{(N)}\}$, где каждый вектор $y^{(i)} \in R^1$ определен на множестве вещественных чисел. В более широкой постановке может рассматриваться несколько зависимых переменных, которые объединяются в вектор $y^{(i)} \in \mathbf{R}^h,\ i = \overline{1,N}$. Полученная функциональная зависимость используется для предсказания ожидаемого значения зависимой переменной для нового набора независимых (входных переменных). Если значение зависимой переменной определяется по отношению к будущему моменту времени, которое присутствует в перечне компонент вектора независимых переменных, то задача регрессии называется *задачей прогнозирования*.

При решении задачи регрессии выполняется построение математической модели, которая определяет характер функциональной связи зависимых и независимых переменных, т. е. проводится оценка параметров получаемой функции регрессии на основе обучающей выборки. Таким образом, основная особенность регрессионного анализа состоит в том, что при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Задача ассоциаций состоит в нахождении значимых связей (ассоциаций) между объектами или событиями. Эти ассоциации представляются в виде определенных правил, которые могут быть использованы для объяснения природы анализируемых процессов и предсказания появления новых событий. Поиск ассоциативных правил является одним из популярных приложений Data Minning [7], при реализации которого ищутся часто встречающиеся наборы объектов среди большого множества таких наборов. По сути, эта задача является частным случаем задачи классификации. Поэтому в дальнейшем, в связи с ограниченным объемом книги, такая задача отдельно рассматриваться не будет, — читатель при этом может обратиться к существующей литературе, например [7, 55].

Помимо основных задач анализа данных, в контексте общей проблемы анализа данных рассматривается еще ряд задач, носящих в определенном смысле вспомогательный характер. К ним относятся задача отбора информативных признаков и задача визуализации.

Отбор информативных признаков и сокращение размерности обычно выполняется на этапе предварительной обработки первичных наблюдений с целью построения описаний адекватных и, в то же время, обозримых описаний объектов.

Визуализация направлена на создание инструментов, позволяющих наглядно отобразить конечный результат выполняемой обработки информации в форме, доступной для понимания человеком. Последняя задача в большей степени носит технологический характер и ее рассмотрение, с учетом ограничений по объему, также выходит за рамки книги.

Все базовые задачи анализа данных делятся [7] на **описательные** (descriptive) и **предсказательные** (predictive). Описательные задачи большее внимание уделяют пониманию обрабатываемых данных. При этом существенное внимание при их решении уделяется обеспечению удобства и прозрачности восприятия получаемых результатов человеком. К таким задачам относятся, прежде всего, задачи кластеризации и поиска ассоциативных правил. Предсказательные задачи обычно осуществляются в два этапа. На первом этапе на основании набора данных с известными результатами строится модель анализируемого процесса или явления. На втором этапе она используется для предсказания результатов на основании

новых наборов данных. При этом, естественно, требуется, чтобы построенные модели работали максимально точно. К такому виду задач относят задачи классификации, регрессии и оценивания. Сюда же можно отнести и задачу поиска ассоциативных правил, если результаты ее решения могут быть использованы для предсказания появления некоторых событий.

Еще одна часто встречающаяся типизация задач обработки информации связана с реализацией обучения, т. е. получения математической модели объектов на основе опытных данных в контексте решаемой задачи. Соответственно возможным ситуациям задачи разделяют на два больших класса, а именно: задачи, реализующие **обучение с учителем** (supervised learning), и задачи, реализующие **обучение без учителя** или **самообучение** (unsupervised learning) [7]. В случае обучения с учителем задача почти всегда решается в два этапа: на первом этапе с использованием обучающей выборки данных строится модель данных (эталонные описания классов, модель регрессии) и на ее основе алгоритм обработки информации, а на втором — проводится тестирование и оценка качества полученного ранее алгоритма обработки информации с использованием тестирующей выборки. При необходимости по результатам тестирования проводится дополнительное обучение. Именно при реализации стратегии обучения без учителя, которое проводится в основном для описательных задач, используется обучающая выборка без использования конкретных априорных сведений о характере данных и их принадлежности.

С точки зрения понимания изложенного ранее материала и, прежде всего, условности введенных понятий и определений, весьма важно подчеркнуть *связь* между различными задачами анализа данных. Следует также отметить, что в системах обработки информации эти задачи часто решаются в комплексе и образуют цепочку функциональных операций, выполняемых как последовательно, так и параллельно, и обеспечивающих достижение общего результата.

Связь задач распознавания (классификации) и кластеризации. Здесь следует отметить, что результаты кластеризации могут быть использованы далее для построения эталонных описаний классов и алгоритма распознавания, предназначенного для отнесения нового объекта, представленного своим образом, к одному из выделенных в ходе кластеризации классов. В таком случае можно говорить о реализации на первом этапе обработки информации обучения без учителя, а на втором этапе — тестирования полученного после обучения алгоритма распознавания. Этот подход часто реализуется в тех ситуациях, когда имеется непомеченная обучающая выборка данных и требуется сформировать в итоге классифицирующее решающее правило.

Связь задач распознавания и оценивания. Такая связь определяется необходимостью реализации процедуры оценивания при проведении обучения классификатора с учителем. При этом, как уже упоминалось, разделяют параметрическое обучение, в рамках которого оцениваются неизвестные параметры эталонных описаний классов (например, неизвестные параметры статистических распределений признаков классов), и непараметрическое обучение, предполагающее проведение оценок эталонных описаний в целом (например, плотностей распределения вероятностей признаков классов). В такой постановке оценивание является неотъемлемой частью процесса обучения и проводится в интересах синтеза алгоритма принятия решений для классификатора.

Связь задач распознавания (классификации) и регрессии. Прежде всего, следует отметить, что обе эти задачи относятся к предсказательным задачам и решаются в два этапа. На первом этапе проводится обучение и формируется модель анализируемых объектов. На втором этапе построенную модель применяют к анализируемым объектам для определения значения зависимой переменной от независимых переменных (входных данных, призна-