УДК 004.93

М.М. Татур, Д.Н. Одинец

О СИСТЕМАТИЗАЦИИ МЕТОДОВ КЛАССИФИКАЦИИ ДАННЫХ И ЗНАНИЙ

С единой методологической позиции систематизируются известные математические методы и алгоритмы классификации. В качестве ключевого признака разделения методов предлагается использовать уровень «интеллектуальности» способа обработки информации. Показывается, что существующие методы используют парадигмы обработки данных либо знаний раздельно, в то время как реальные задачи распознавания могут включать в себя оба подхода одновременно. Обосновывается вывод о необходимости развития моделей классификации в направлении унификации и гибридизации.

Введение

В современной литературе к математическим методам классификации авторы относят достаточно широкий список результатов. Перечислим наиболее известные из них:

– байесовская стратегия принятия решений [1–4];
– метод *k*-ближайших соседей [3, 4];
– метод сравнения с эталоном по минимуму расстояний [3–5];
– метод сравнения с эталоном по максимуму функции корреляции [6–8];
– метод опорных векторов (support vector machine) [4–9];
– различные нейросетевые методы [4, 10, 11];
– методы обучения, включая итеративные, группового учета аргументов, генетический алгоритм [10–12];
– деревья решений [1, 3, 4];
– логический и нечеткий вывод [10–12];
– методы структурного и синтаксического распознавания [1, 13].

Как правило, указанные результаты приводятся в виде абстрактных математических методов, без структурирования и аналитики. Наиболее часто приводятся результаты применения одного либо нескольких методов для решения конкретных задач, при этом сравнительные результаты эффективности методов не претендуют на фундаментальный характер. В ряде случаев, в том числе и в указанных выше ссылках на использованные источники, нет четкого терминологического разделения процессов классификации, принятия решения и обучения. Для устранения терминологических неточностей введем следующую систему терминов и ограничений.

1. Основные термины и ограничения

Теория классификации, как и любая теория, представлена системой терминов, определений, положений, методов и теорем. В настоящей работе предлагается авторская систематизация наиболее значимых результатов в данной области и определяется направление для развития новых методов классификации. Чтобы выполнить данную часть работ, необходимо сделать следующие терминологические уточнения, допущения и ввести некоторые ограничения [14].

Классификатор – это абстрактный автомат, принимающий решение об отнесении образа к одному из классов из некоторого перечня. Образ (вектор информативных признаков) – это совокупность значений некоторых характеристик объекта классификации, перечисленных в определенном порядке и интерпретируемых как координаты точки в *N*-мерном пространстве, где *N* зависит от числа характеристик объекта. Конечный результат классификации определяется критерием принятия решения. Примерами простейших критериев являются: максимум достоверности, минимум ошибок первого либо второго родов, минимум потерь в результате неправильно принятого решения и т. п. Все критерии в конечном итоге сводятся либо к выбору оптимального уровня порога для четкого способа принятия решения, либо к выбору параметров функции нечеткого принятия решения.

Идентификация представляет собой частный случай классификации, когда принимается решение об отнесении (либо нет) вектора информативных признаков к заданному классу. Иногда идентификацию называют двухклассовой задачей распознавания. Поскольку для задач идентификации удобно получать количественные оценки эффективности (достоверности принятия решений), интерпретировать и оптимизировать процесс классификации, то зачастую многоклассовые задачи сводятся к совокупности задач идентификации для множества классов. В литературе часто теоремы и утверждения доказываются, иллюстрируются для задач идентификации, а затем обобщаются для многоклассовых задач. В этом случае принятие решения может осуществляться на различных уровнях, с различными критериями.

В дальнейшем математические модели классификаторов будут рассматриваться применительно к задачам идентификации. Многоклассовые задачи будут представлены совокупностью задач идентификации, или элементарными классификаторами.

Методы классификации часто рассматривают в контексте обучения классификатора, т. е. его настройки на некоторой совокупности векторов, с известными откликами (обучающая последовательность). В литературе термин «обучение» часто сводится к выбору структуры и обучению некоторой нейронной сети. Между тем в более широком смысле, применительно к другим методам классификации, обучение может трактоваться совершенно иначе, например как формирование эталона, поиск доверительных интервалов, выявление зависимостей между данными, оптимизация критерия принятия решения и т. п.

Исследование алгоритмов обучения, способов формирования обучающих последовательностей и сравнительного анализа эффективности различных классификаторов на тестовых данных представляется одним из важных этапов разработки систем распознавания, однако рассмотрение этих вопросов не является целью настоящей работы и будет предметом обсуждения в последующих публикациях.

2. Методы, основанные на обработке данных

Под термином «данные» в контексте настоящей работы понимаются результаты измерения некоторых физических величин: веса, размера, температуры, частоты, скорости и т. п. – либо вычисления математических величин: математического ожидания, дисперсии, производной, функции корреляции и т. п. Данные могут быть как однородными: логическими, количественными, номинальными (цвет, сорт и т. п.), так и иметь сложную структуру. Когда речь идет о методах классификации, основанных на обработке данных, полагается, что зависимость между данными неизвестна.

Классификация образов посредством измерения расстояния между векторами информативных признаков является одним из первых, интуитивно-понятных подходов. Два вектора x^* и x имеют расстояние (различие) $d(x^*, x)$ в некоторой метрике. В теории распознавания известны и применяются следующие меры (метрики): хэммингово расстояние, функции корреляции, направляющие косинусы, евклидово расстояние, меры сходства Минковского, мера сходства Танимота, взвешенные меры сходства, операция сравнения с использованием нечеткой логики, вариационные методы, процедура динамической свертки, левенштейново расстояние, сравнение по неизменным признакам и др.

Пусть один из векторов (например, x^*) является входным вектором, а второй вектор может являться одним из векторов обучающей последовательности x_i либо некоторым усредненным вектором эталонного образа x^o. В зависимости от этого различают методы ближайшего соседа (k ближайших соседей), метод сравнения с эталоном и различные их модификации.

2.1. Метод ближайших соседей

Пусть $X_p = \{x_1, x_2, \ldots, x_p\}$ – множество векторов обучающей последовательности и принадлежность каждого из них тому или иному классу достоверно известна. Для всех векторов обучающей последовательности определяют расстояния в некоторой метрике и выбирают x_i с минимальным $d(x^*, x_i)$. Правило классификации $x^* \notin X_p$ состоит в том, что x^* относят к тому классу, которому принадлежит x_i. Очевидно, что такое отнесение носит случайный характер. Чтобы повысить вероятность правильного решения, задается число k (обычно от 3 до 10) и для

образа, который нужно классифицировать, находится k ближайших (по расстоянию) соседей, а класс неизвестного образа определяется «большинством голосов».

2.2. Метод сравнения с эталоном

Для каждого класса по обучающей выборке строится эталон x^{o}, имеющий следующие значения x_i^{0} в N-мерном пространстве признаков:

$$x^{o} = \{x_1^{0}, x_2^{0}, \ldots, x_N^{0}\},$$

где $x_i^{0} = \frac{1}{K}\sum_{k=1}^{K} x_{ik}$; K – количество векторов данного класса в обучающей выборке; i – номер признака.

По существу, эталон – это усредненный по обучающей выборке абстрактный образ. Абстрактным его называют потому, что он может не совпадать ни с одним вектором обучающей выборки. Между эталоном и меткой класса устанавливается однозначная связь. Классификация осуществляется по следующему алгоритму. На вход поступает образ x^*, принадлежность которого к тому или иному классу неизвестна. От этого вектора измеряются расстояния до всех эталонов $d(x^*, x^0)$, и система относит текущий вектор к тому классу, расстояние до эталона которого минимально. Расстояние измеряется в той метрике, которая введена для решения конкретной задачи классификации. Принято считать, что эффективность классификации данных с помощью функции расстояния приемлема лишь в ограниченных случаях. На практике эти методы неконкурентоспособны с методами, основанными на разделении классов.

Альтернативой методам, основанным на измерении расстояния между образами, является группа методов, основанных на *разделении образов в пространстве признаков* некоторой гиперплоскостью $D(X)$:

$$D(X) = \sum_{i=1}^{N} a_i x_i + a_0 .$$

В простейшем случае, когда образы могут быть разделены границей, описываемой линейной функцией $D(X)$, их называют линейно разделимыми. Наглядно это можно проиллюстрировать для случая распознавания двух классов, когда число признаков $N = 2$. Тогда разделяющая гиперплоскость вырождается в линию на плоскости, а функцию разделения иногда называют линейным решающим правилом при следующих условиях:

$D(X) > 0$, где X – векторы первого класса;

$D(X) < 0$, где X – векторы второго класса.

2.3. Нейронные сети

Существуют различные методы построения (реализации) линейных решающих правил. Один из наиболее известных, предложенный в 1950-х гг. Розенблатом, был назван персептроном и открыл целое направление – искусственные нейронные сети (ИНС). В основу теории ИНС положена модель формального нейрона. В классическом варианте входные сигналы нейрона соответствуют вектору $X = \{x_1, x_2, \ldots, x_N\}$, каждый вход x_i умножается на весовой коэффициент w_i (i = 1, 2, …, N), соответствующий «силе» синапсической связи, и все произведения суммируются. Суммарное возбуждение пропускается через активационную функцию $F(*)$, в результате чего определяется выходной сигнал:

$$Y = F\left(\sum_{i=1}^{N} w_i \cdot x_i - \theta\right),$$

где θ – смещение порога.

Как правило, активационная функция является нелинейной «сжимающей», т. е. нормирует диапазон выходных значений Y в интервале [0,1]. Число входов нейрона N определяется размерностью N-мерного пространства, в котором входные сигналы могут быть представлены точками или областями из близко расположенных точек (для нечетко задаваемых образов). Один нейрон описывает гиперплоскость в таком N-мерном гиперпространстве и разделяет его на две непересекающиеся и невложенные гиперобласти. Например, при $N = 2$ и пороговой функции активации с помощью такого нейрона может быть решена простейшая задача распознавания двух классов образов (т. е. задача идентификации):

$$Y = \begin{cases} 1, \text{ если } x_1 w_1 + x_2 w_2 \geq \theta; \\ 0, \text{ если иначе.} \end{cases}$$

Существует множество вариаций и дополнений к стандартной структуре нейрона, разработка которых обусловлена поиском оптимальных нейросетевых решений в конкретных прикладных задачах [5, 11, 12]. Функция активации нейрона также может иметь различный вид, определяемый функциональным назначением нейрона в составе ИНС.

Описания образов в реальных задачах распознавания, как правило, образуют в пространстве признаков сильно пересекающиеся или вложенные гиперобласти, разделить которые можно только проведя нелинейную гиперповерхность между ними. Такая гиперповерхность аппроксимируется с помощью полинома

$$D(X) = \sum_{k=1}^{N3} c_k \left(\sum_{j=1}^{N2} b_j \left(\sum_{i=1}^{N1} a_i x_i \right)_j \right)_k ,$$

где i, j, k – количество аппроксимирующих отрезков в каждой из плоскостей; a, b, c – постоянные коэффициенты.

С позиций теории нейронных сетей это означает переход от отдельного нейрона к совокупности определенным образом взаимосвязанных нейронов – многослойной нейронной сети. Нейросетевой подход к построению нелинейных разделяющих поверхностей обычно реализуется при помощи композиции более простых с вычислительной точки зрения – линейных – процедур. Так, многослойная нейронная сеть прямого распространения (многослойный персептрон) состоит из входного слоя, нескольких скрытых слоев и выходного слоя нейронов [10]. Количество нейронов во входном слое соответствует размерности вектора информативных признаков. Нейроны входного слоя служат лишь для распределения входных сигналов между нейронами скрытого слоя и не выполняют каких-либо вычислений. Каждый из нейронов скрытых слоев осуществляет нелинейное преобразование сигналов, поступающих на его вход; чем больше нейронов и чем больше слоев, тем более сложные кусочно-линейные разделяющие гиперповерхности может реализовать ИНС, тем более сильно пересекающиеся образы она сможет различать, тем больше емкость ее ассоциативной памяти. Количество нейронов в выходном слое определяет число распознаваемых классов.

Существуют и другие разновидности методов, в той или иной мере реализующие идею разделения образов. К ним можно отнести метод опорных векторов (Support Vector Machine), метод потенциальных функций и др.

3. Методы, основанные на обработке знаний

Под термином «знание» в контексте настоящей работы понимаются элементарные зависимости между данными и конечным результатом, которые устанавливаются на этапе построения классификатора либо отыскиваются (извлекаются) на стадии обучения. Далее эти зависимости могут компилироваться и использоваться для принятия решений различными методами, такими, как деревья решений, логический и нечеткий выводы. В более сложных проявлениях знания выступают как иерархическая структура, оперирующая библиотеками примитивов более низких уровней. На каждом уровне могут независимо использоваться свои методы локального распознавания. К таким методам можно отнести цепное кодирование, анализ грамматик, семантические сети и т. п.

3.1. Деревья решений

Известен целый ряд результатов, связанных с построением деревьев (своего рода обучением классификатора) в различных прикладных областях, например в медицинской и технической диагностике. Обычно деревья решений применяют для решения многоклассовых задач, но с тем же успехом этот метод можно применять и для задач идентификации.

Основное предназначение метода состоит в том, чтобы обеспечить наглядность и сократить перебор при реализации функции принятия решений. В качестве примеров (рис. 1, *а*) приведен частный случай дерева для двух признаков $\{x_1, x_2\}$ и четырех классов $\{y_1, y_2, y_3, y_4\}$, а также для идентификации по признакам $\{x_1 \div x_7\}$ (рис. 1, *б*). В данных примерах в каждой вершине графа осуществляется проверка условия по одному признаку. Сокращение времени (объема) вычислений будет происходить за счет сокращения числа анализируемых признаков.

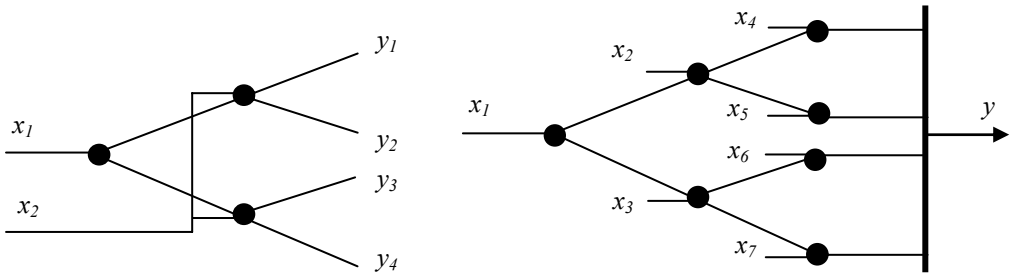

Рис. 1. Граф-схемы вычислений при реализации деревьев решений: *а)* в задаче классификации четырех классов; *б)* задаче идентификации по восьми информативным признакам

Хорошо известно, что существует множество вариантов деревьев решений с различными критериями разделения и топологиями. Объединяет их лишь один аспект: все они функционируют по последовательному принципу, а для вычислительного процесса достаточно «пройти» только по одной ветви. Необходимыми условиями для применения данного метода являются:

– наличие объективной зависимости разделения (расщепления) групп классов от значения одного (либо нескольких) признаков;

– знание этой зависимости.

3.2. Логический вывод

Вывод – это процесс рассуждения, в ходе которого осуществляется переход от некоторых исходных суждений (предпосылок) к новым суждениям – заключениям. Правила преобразования исходной системы предпосылок в систему заключений называются правилами логического вывода, или правилами продукций. Если вид предпосылок и заключений указан явно, вывод называется прямым. Основными критериями качества правил логического вывода являются их полнота и непротиворечивость [15, 16].

Для решения задач идентификации будем использовать элементы теории логического вывода путем интуитивной записи правил продукций типа «ЕСЛИ – ТО». Такие правила состоят из ряда начальных условий-предпосылок и заключений, например:

$$\text{ЕСЛИ} \quad x_1 \ \text{есть}\ A_{11} \ \text{И} \ x_2 \ \text{есть}\ A_{12} \ \ldots \ x_n \ \text{есть}\ A_{1n} \quad \text{ТО} \ y \ \text{есть} \ C_1;$$

$$\ldots$$

$$\text{ЕСЛИ} \quad x_1 \ \text{есть}\ A_{k1} \ \text{И} \ x_2 \ \text{есть}\ A_{k2} \ \ldots \ x_n \ \text{есть}\ A_{kn} \quad \text{ТО} \ y \ \text{есть} \ C_k,$$

где $x_1, x_2, \ldots. x_n$ – входной n-мерный вектор информативных признаков; $A_{11} \ \ldots \ A_{kn}$ – значения соответствующих эталонов; y – выходная переменная; $C_1 \ \ldots \ C_k$ – идентификаторы классов.

В данном примере каждому классу соответствует одно правило вывода в виде конъюнктивного терма, в котором присутствуют все информативные признаки. В общем случае правило

вывода для каждого класса может содержать k конъюнктивных термов, объединенных по ИЛИ, тогда для задачи идентификации правило вывода примет вид

$$\text{ЕСЛИ } x_1 \text{ есть } A_{11} \text{ И } x_2 \text{ есть } A_{12} \dots x_n \text{ есть } A_{1n} \text{ ИЛИ}$$
$$\dots$$
$$\text{ЕСЛИ } x_1 \text{ есть } A_{k1} \text{ И } x_2 \text{ есть } A_{k2} \dots x_n \text{ есть } A_{kn} \text{ ТО}$$
$$y \text{ есть ДА, ИНАЧЕ – НЕТ.}$$

Необходимые условия применения данного метода:

– наличие объективной логической зависимости заключения (решения) от состава и значений информативных признаков;

– знание этой зависимости.

Таким образом, как для деревьев решений, так и для логического вывода необходимы знания правил разделения либо логической зависимости. Получение этой зависимости можно рассматривать как извлечение «элементарного знания» из массива данных, а все последующие знания – знания «более высоких порядков» – в конечном итоге сводятся к композиции элементарных знаний.

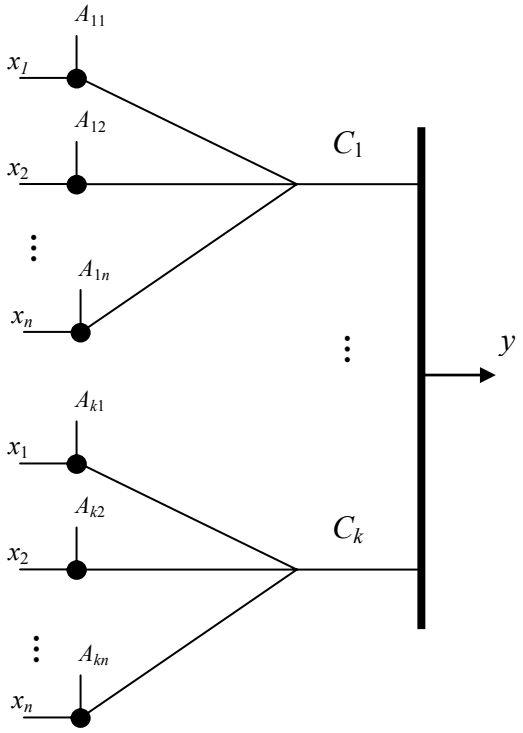

Рис. 2. Граф-схема вычислений при реализации общего случая логического вывода

По сравнению с деревьями решений метод логического вывода в общем случае предполагает выполнение вычислений по всем ветвям графа (рис. 2). Таким образом, сокращение объема вычислений общего времени не происходит, а сокращение времени вычислений может достигаться только за счет реализации на параллельной архитектуре компьютера. Однако с функциональной точки зрения дерево решений может быть представлено системой правил логического вывода, следовательно, может рассматриваться как частный случай последнего. Деревья решений и логический вывод могут быть реализованы в формате четкого принятия решения с применением библиотеки арифметической операции компарирования и логических операций И – ИЛИ. Деревья решений по своей сути предполагают четкость в трактовке правил разделения, поскольку за цикл реализуется лишь одна ветвь графа, а вот логический вывод, напротив, может реализоваться в нечетком формате принятия решений.

3.3. Нечеткий вывод

Классическая логика оперирует значениями «истина» и «ложь», однако этими двумя значениями довольно сложно, а порой и вовсе невозможно представить большое разнообразие реальных задач классификации и распознавания. Если в описании величин присутствует неточность, применяется аппарат теории вероятностей. Тогда неопределенность состоит в принадлежности или непринадлежности некоторого объекта либо события к обычному (четкому) множеству.

В качестве альтернативы была предложена теория нечетких (размытых) множеств, авторство которой принадлежит американскому математику Лотфи Заде [17]. Теория была направлена на преодоление проблем, связанных с представлением неточных понятий при анализе и моделировании систем. Основным отличием нечеткой логики от классической, как следует из названия, является наличие помимо двух предельных состояний «истина» и «ложь» неограниченного числа промежуточных. Такое (нечеткое) представление данных и ограничений в виде функций принадлежности дает возможность получать устойчивые решения в условиях погрешности информации и нечеткости исследуемых процессов с указанием степени снижения качества.

В теории нечеткой логики вводятся расширения базовых операций логического умножения, сложения и отрицания операциями минимума, максимума и дополнения:

$$a \wedge b = \min\{a,b\};$$
$$a \vee b = \max\{a,b\};$$
$$\overline{a} = 1 - a.$$

Легко заметить, что при использовании только двух состояний («ложь» – 0 и «истина» – 1) результаты вычислений в нечеткой логике сводятся к результатам в классической бинарной логике.

Основой для проведения операции нечеткого логического вывода является база правил, содержащая нечеткие высказывания в форме «ЕСЛИ – ТО» и функции принадлежности для соответствующих лингвистических термов.

Пусть в базе правил имеется k правил вида

$$\text{ЕСЛИ } x_1 \text{ есть } A_{11} \text{ И } x_2 \text{ есть } A_{12} \ldots x_n \text{ есть } A_{1n} \text{ ИЛИ}$$
$$\text{ЕСЛИ } x_1 \text{ есть } A_{k1} \text{ И } x_2 \text{ есть } A_{k2} \quad x_n \text{ есть } A_{kn} \text{ ТО}$$
$$y \text{ есть } B, \text{ ИНАЧЕ – НЕТ},$$

где x_1, x_2, …. x_n – входной n-мерный вектор информативных признаков; A_{11} … A_{kn} и B – значения лингвистических переменных, заданных функциями принадлежностей.

В общем случае механизм логического вывода включает следующие этапы: введение нечеткости (*фаззификация*), собственно нечеткий вывод и *дефаззификация*. Известны модели нечеткого вывода Мамдани, Сугено, Ларсена, Цукамото, которые различаются видами используемых правил, логических операций и способами дефаззификации [15]. В последние годы наметилась тенденция к гибридизации методов интеллектуальной обработки информации. В результате объединения нескольких технологий искусственного интеллекта появился специальный термин «мягкие вычисления» (*soft computing*), который ввел Л. Заде в 1994 г. В настоящее время мягкие вычисления объединяют такие области, как нечеткая логика, искусственные нейронные сети, вероятностные рассуждения и эволюционные алгоритмы. Они дополняют друг друга и используются в различных комбинациях для создания гибридных интеллектуальных систем [18]. Например, нечеткие нейронные сети (*fuzzy-neural networks*) осуществляют выводы на основе аппарата нечеткой логики, однако параметры функций принадлежности настраиваются с использованием алгоритмов обучения нейронных сетей [19, 20]. Для подбора параметров таких сетей применим метод обратного распространения ошибки, изначально предложенный для обучения многослойного персептрона. Алгоритмы настройки нечетких систем на решение конкретной задачи относительно трудоемки и сложны по сравнению с алгоритмами обучения нейронных сетей, потому что, как правило, состоят из двух задач:

– генерации лингвистических правил (извлечения знаний);

– настройки и корректировки функций принадлежности.

Первая задача относится к задачам переборного типа, вторая – оптимизации в непрерывных пространствах. При этом возникает определенное противоречие: для генерации нечетких правил необходимы функции принадлежности, а для настройки функций принадлежностей – правила нечеткого вывода. В процессе настройки необходимо обеспечивать полноту и непротиворечивость правил нечеткого вывода. Как самостоятельный подход в обучении нечетких систем применяется генетический алгоритм.

3.4. Методы структурного, синтаксического и семантического распознавания

Было замечено, что, оперируя ограниченным числом примитивов (непроизводных элементов), можно описывать больше разнообразных объектов. Для того чтобы выполнить такое описание, наряду с выделением примитивов должны вводиться правила комбинирования, или композиции. При построении описания какого-либо объекта непроизводные элементы объединяются в цепочки (предложения) по определенному набору правил. В этом структурные методы аналогичны процессу построения предложений естественного языка. В результате связей между непроизводными элементами (структурными признаками) образуется объект. Это аналогично построению предложений языка путем соединения слов, которые состоят из букв. Поэтому структурные признаки носят еще название лингвистических или синтаксических. Например, если задана операция конкатенации, то каждый объект представляется цепочкой примыкающих непроизводных элементов. Решение о том, является ли представление объекта синтаксически правильным (т. е. принадлежит ли он к классу образов, описываемых данным синтаксисом или данной грамматикой), принимается блоком синтаксического анализа (классификатором) или блоком грамматического разбора [13]. Очевидно, что простейшим методом распознавания является сравнение с эталоном. По ходу синтаксического анализа или грамматического разбора этот блок может давать полное синтаксическое описание объекта в терминах грамматических единиц или дерева грамматического разбора, если представление объекта синтаксически правильно. В противном случае объект либо исключают из рассмотрения, либо анализируют на основе других заданных грамматик.

При наличии шума и искажений два или несколько разных образов могут быть представлены как весьма похожие объекты. В рамках синтаксического подхода это означает, что один образ имеет два или несколько разных структурных описаний. В лингвистической теории такая ситуация называется неоднозначностью. Эту неоднозначность структурных описаний нельзя разрешить обычными процедурами синтаксического анализа. Необходимое различение можно провести, применяя вероятностные методы, в которых учитывается статистическая информация об искажениях и шуме, либо используя априорные сведения, семантику контекста. Семантические сети являются еще более высоким уровнем представления знаний и представляют самостоятельное научное направление, которому соответствуют свои специфические методы отождествления информации и принятия решений.

4. Систематизация рассмотренных математических методов классификации

Несмотря на широкое многообразие математических методов классификации и глубокие исследования в отдельных узких направлениях, аналитические работы, в которых предлагаются научно обоснованные подходы к систематизации накопленных знаний в данной области, встречаются крайне редко. На основании проведенного анализа источников [1–20] можно утверждать, что общепринятого подхода к систематизации известных результатов в области математических методов классификации не существует. В основном результаты представлены разрозненно, без должного обобщения, что зачастую ведет к путанице и эклектике, повторению схожих результатов под другими, порой броскими названиями, к затруднению корректной оценки действительно новых результатов. Такая ситуация отрицательно сказывается на развитии теории классификации в целом.

Идея предлагаемого подхода к систематизации известных результатов состоит в том, чтобы разместить методы в направлении роста уровня «интеллектуальности». Так, основные методы, обзор которых выполнен выше, разделены на две группы (рис. 3):

1) основанные на обработке данных;

2) основанные на обработке знаний.

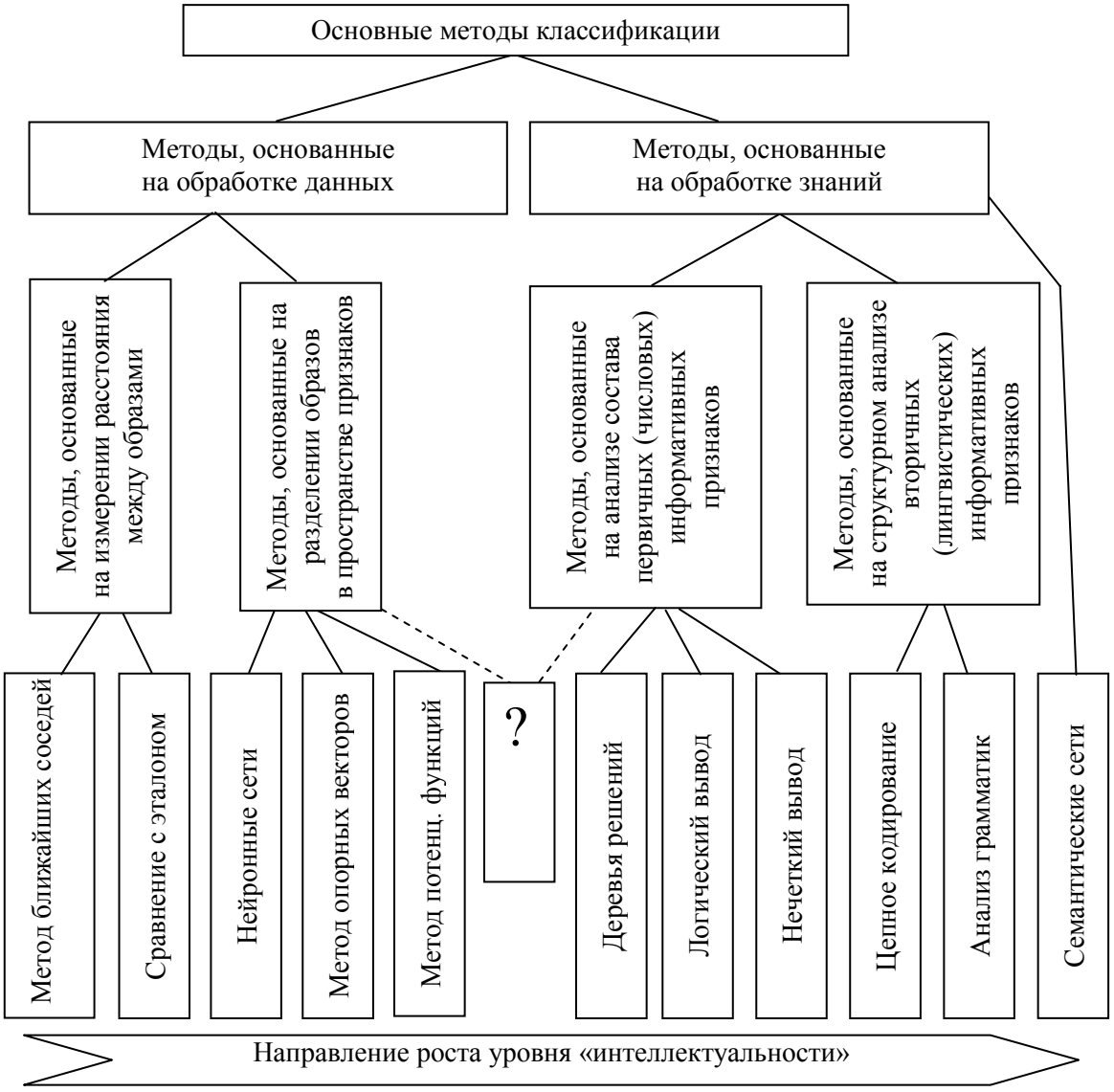

Рис. 3. Систематизация основных методов классификации

Первую группу методов объединяет то, что информативные признаки между собой никак не связаны, т. е. выступают как первичные данные (отсюда и название группы). В свою очередь, эти результаты включают методы, основанные:

– на измерении расстояния между образами;

– на разделении образов в многомерном пространстве признаков.

Вторая группа методов классификации включает результаты, качественно отличающиеся от методов первой группы. Было замечено, что многие информативные признаки логически связаны между собой. Такую связь можно рассматривать как «*элементарное знание*». Если эти элементарные знания найти (извлечь из огромного потока данных), объем дальнейших вычислений при классификации можно значительно сократить, а эффективность классификации повысить.

К первой подгруппе методов второй группы можно отнести различные деревья решений, логический и нечеткий выводы. Общим для таких методов является то, что информативные признаки представляют собой численные значения. Это в некоторой степени сближает их с первой группой методов, основанных на обработке данных. В данной подгруппе играет роль (и анализируется) только состав признаков.

Вторая подгруппа методов характеризуется тем, что информативные признаки не являются просто числами, а представляют собой некоторые структурные примитивы, прошедшие этап предварительного распознавания и выделенные для дальнейшего участие в этапах интел-

лектуальной обработки. Здесь важен не только их состав, но и взаимное расположение (например, во времени или в пространстве). С помощью этих методов можно в простых формах записывать знания более высоких порядков и соответственно анализировать и принимать классификационные решения. К этой подгруппе можно отнести достаточно большой перечень методов: от совершенно очевидных (например, цепное кодирование), до сложных (например, семантические сети и структурное распознавание).

В следующей публикации будет предложена математическая модель классификатора, которая продемонстрирует плавную трансформацию от «обработки данных» к «обработке элементарных знаний». Местоположение такой модели в предложенном систематизаторе обозначено на рис. 3 прямоугольником со знаком вопроса.

Заключение

Аналитический обзор математических методов классификации позволяет сделать следующие выводы:

1. Методы классификации существуют и применяются разрозненно. Попытки гибридизации, объединения методов, как правило, носят эвристический эклектичный характер.

2. Методы классификации имеют явно выраженную «интеллектуальную» иерархию – от простейших до сложных. На этих принципах предложено оригинальное структурирование известных результатов, что согласуется с парадигмами психологии: восприятие – осознание – рассуждение.

3. В реальных задачах классификации нет жесткого разграничения между обработкой данных и знаний. Часто при распознавании фотопортретов, отпечатков пальцев, сетчатки глаза и других сложных объектов требуется совместная обработка данных и знаний. Однако эффективных моделей, которые бы отражали диалектический переход от одной группы методов к другой или вырождение сложных методов в более простые, нет. Особенно этот «разрыв» проявляется в различии методов, основанных на обработке данных и знаний.

Список литературы

1. Дуда, Р. Распознавание образов и анализ сцен / Р. Дуда, П. Харт. – М. : Мир, 1976. – 512 с.

2. Белозерский, Л.А. Введение в системы автоматического распознавания / Л.А. Белозерский. – Киев : Наукова думка, 2005. – 434 с.

3. Васильев, В.И. Распознающие системы / В.И. Васильев. – Киев : Наукова думка, 1983. – 422 с.

4. Чубукова, И.А. Data Mining / И.А. Чубукова. – М. : БИНОМ, 2006. – 382 с.

5. Кохонен, Т. Самоорганизующиеся карты / Т. Кохонен. – М. : БИНОМ, 2008. – 655 с.

6. Василенко, Г.И. Оптико-электронные корреляционные методы и средства распознавания изображений / Г.И. Василенко, И.С. Гибин, О.И. Потатуркин // Радиоэлектроника. – 1990. – № 8. – С. 15–27.

7. Zholtikov, R.R. Some models of raster correlators of binary images / R.R. Zholtikov, M.M. Tatur // Int. Scientific Journal of Computing. – 2004. – Vol. 3. – P. 46–49.

8. Устройство корреляционного распознавания бинарных образов : пат. 1748 Респ. Беларусь, МПК7, G 06K 9/00 / М.М. Татур, Р.Р. Жолтиков ; заявитель М.М. Татур. – № u 200402333 ; заявл. 2004.10.01 ; опубл. 30.03.2005 // Афіцыйны бюл. / Нац. цэнтр інтэлектуал. уласнасці. – 2005. – № 1. – С. 125.

9. SVM – Support Vector Machines, USA [Electronic resource]. – Mode of access : http://www.support-vector-machines.org. – Date of access : 10.06.2010.

10. Головко, В.А. Нейроинтеллект: теория и применение. Кн. 1. Организация и обучение нейронных сетей с прямыми и обратными связями / В.А. Головко. – Брест : Университетское, 1999. – 260 с.

11. Уоссермен, Ф. Нейрокомпьютерная техника. Теория и практика / Ф. Уоссермен. – М. : Мир, 1992. – 240с.

12. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилинский, А. Рутковский. – М. : Горячая линия – Телеком, 2006. – 383с.

13. Фу, К. Структурные методы в распознавании образов / К. Фу. – М. : Мир, 1977. – 320 с.

14. Татур, М.М. Формальные и неформальные аспекты в разработке систем распознавания / М.М. Татур // Искусственный интеллект. – 2007. – № 3. – С. 333–343.

15. Леоненков, А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH / А.В. Леоненков. – СПб. : БХВ–Петербург, 2003. – 720 с.

16. Штовба, С.Д. Классификация объектов на основе нечеткого логического вывода / С.Д. Штовба // Методы. Алгоритмы. Программы. – 2004. – № 1. – С. 68–69.

17. Заде, Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л.А. Заде. – М. : Мир, 1976. – 168 с.

18. Trends in practical Applications of Agent and Multiagent Systems // Advances in Intelligent and Soft Computing. – Berlin, Germany : Springer-Verlag Berlin Heidelberg, 2010. – Vol. 71. – 733 p.

19. Новоселова, Н.А. Построение нечеткой модели для решения задач классификации / Н.А. Новоселова // Информатика. – 2006. – № 3 (11). – С. 5–14.

20. Ярушкина, Н.Г. Нечеткие гибридные системы. Теория и практика / Н.Г. Ярушкина. – М. : Физматлит, 2007. – 208 с.

Поступила 24.05.10

Белорусский государственный университет
информатики и радиоэлектроники,
Минск, П. Бровки, 6
e-mail: tatur@bsuir.by

M.M. Tatur, D.N. Adzinets

ON SYSTEMATIZATION OF DATA AND KNOWLEDGE CLASSIFICATION METHODS

The original author's approach to systematization of the most significant methods and algorithms of classification is presented. As a key feature for separating existing methods, the level of «intelligence» of information processing is employed. It is shown that the known methods make use paradigms of data and knowledge processing separately whereas the applied recognition tasks may include both approaches taken simultaneously. The conclusion about the necessity of the development of classification models towards corresponding unification and hybridization has been drawn.