УДК 519.95

ЭМПИРИЧЕСКОЕ ОБОБЩЕНИЕ И РАСПОЗНАВАНИЕ: КЛАССЫ ЗАДАЧ, КЛАССЫ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ И ПРИМЕНИМОСТЬ ТЕОРИЙ. ЧАСТЬ I

© Донской В.И.

Таврический национальный университет им. В.И. Вернадского
факультет математики и информатики
пр-т Вернадского, 4, г. Симферополь, 95007, Украина
e-mail: *donskoy@tnu.crimea.ua*

Abstract. Classification of pattern recognition problems is offered. This classification is founded on the basic properties of pattern recognition problems. It is shown, a choice of methods of decisions must be coordinated with features of classes of pattern recognition problems.

Введение

Современное состояние распознавания образов как науки характеризуется появлением вполне обоснованных теорий, иногда базирующихся на существенно различающихся исходных предположениях. Широко известны алгебраическая теория распознающих и классифицирующих алгоритмов Ю. И. Журавлева [4], статистическая теория распознавания В. Н. Вапника и А. Я. Червоненкиса [1,2], статистическая параметрическая теория, базирующаяся на байесовском подходе и ведущая свое начало от работ Р. Фишера, структурно-лингвистические теории, MDL, другие теории и их модификации. Закономерно возникает вопрос о применимости каждой из рассматриваемых теорий к различным задачам и выделении в этой связи специфических классов задач. С указанным вопросом также связаны выбор моделей распознавания, методов обучения и их обоснование. Выбор подхода к решению конкретной задачи распознавания и его обоснование – нетривиальная проблема. Но она, как правило, остается в тени; усилия исследователей направлены на создание алгоритмов обучения и оценивание вероятности ошибок распознавания [1-4]. Существен и педагогический аспект рассматриваемой в настоящей работе проблемы: преподавание предмета «Математические методы распознавания образов». Здесь неизбежно возникает вопрос сравнительного анализа областей применимости различных теорий.

Приведенные соображения, возникшие в процессе научной и педагогической деятельности в области распознавания образов, побудили автора к написанию настоящей статьи. Цель работы – изложить и обосновать подход к определению областей применимости (или неприменимости) основных теорий обучения и распознавания. Соответствующие области представляют собой классы задач распознавания, определяемые общностью их основных свойств. Будем также называть такие классы задач распознавания семействами, если слово «класс» будет использоваться в контексте для обозначения множества объектов генеральной совокупности.

1. Концептуальная схема построения правил распознавания и классификация решаемых задач

На рис. 1 приведена схема, представляющая связь объектов и процессов, происходящих при построении решающих правил распознавания.На этой схеме указаны некоторые свойства рассматриваемых объектов. В соответствии с концептуальной схемой выделен круг признаков, используя которые можно описать классы решаемых задач распознавания.

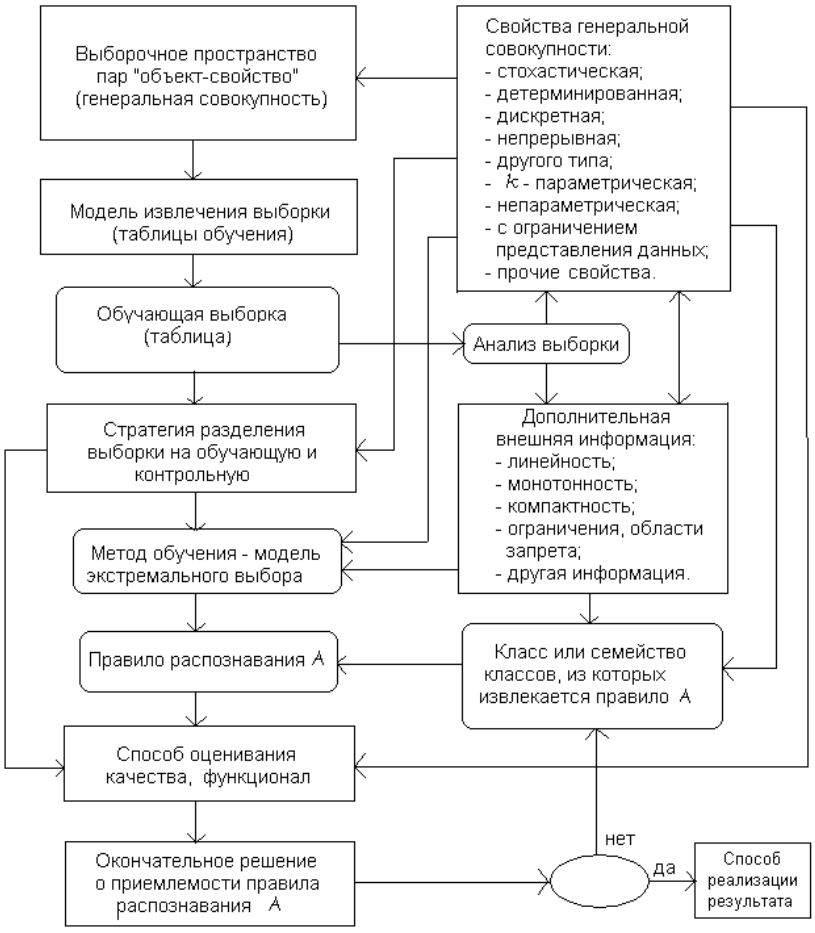

Рис. 1. Связь основных объектов и процессов, происходящих при выборе решающих правил распознавания

Выборочное пространство $\mathbf{X}$ состоит из объектов $X = (x_1, \ldots, x_n)$, называемых допустимыми, компоненты которых (переменные-признаки) принимают значения из множеств D_i, $i = 1, \ldots, n$. Эти множества могут быть непрерывными,

дискретными, разнотипными. Полагается существующим некоторый набор основных свойств (предикатов) $\omega_j : D_1 \times \cdots \times D_n \longrightarrow \{0;1\}$, $j = 1,\ldots,s$. Множества $K_j = \{X \in \mathbf{X} : \omega_j(X) = 1\}$ называют классами; $\tilde{\omega} = (\omega_1(X),\ldots,\omega_s(X))$ – двоичный вектор, определяющий принадлежность объекта классам. Значения этого вектора можно считать номерами классов (при пересекающихся классах-номерами комбинаций классов). Выборочное пространство $\mathbf{X}$ является генеральной совокупностью объектов, из которой извлекается конечное подмножество объектов $\{X_1,\ldots,X_l\} = \tilde{X}_l$ – выборка, которая вместе с полученными некоторым способом значениями $\tilde{\omega}(X_1),\ldots,\tilde{\omega}(X_l)$ образует таблицу обучения. Таблица обучения – это совокупность пар $\{(X_q,\tilde{\omega}(X_q)), q = 1,\ldots,l\}$. Выборочное пространство обладает набором свойств, которые отражаются в таблице обучения.

Вообще говоря, не исключено, что таблица обучения может иметь пропуски в данных и ошибки – как в значениях признаков, так и в значениях принадлежности классам. Задача обучения распознаванию состоит в нахождении по таблице обучения решающей функции, позволяющей правильно (или приближенно, но как можно более точно) находить для любого объекта $\mathbf{X}$ из генеральной совокупности значение номера класса $\tilde{\omega}(X)$. Будем, обобщая, называть задачей распознавания задачу обучения вместе с задачей использования найденной решающей функции для определения номера класса $\tilde{\omega}(X)$ произвольного $X \in \mathbf{X}$.

В соответствии с концептуальной схемой, основные свойства задач распознавания сведены в представленную ниже таблицу. В стохастических задачах все данные в таблице обучения являются случайными величинами, извлеченными из генеральной совокупности, вообще говоря, с неизвестными законами распределения. В некоторых случаях эти законы полностью или с точностью до параметров априорно известны. В детерминированных задачах как объекты в генеральной совокупности, так и вся информация в таблицах обучения достоверны, но возможны пропуски и/или ошибки в данных, связанные с процессом их извлечения. В недетерминированных задачах часть данных не определена, и нет никакой дополнительной информации об их возможных значениях.

Модель извлечения обучающей выборки определяет схему её выбора из генеральной совокупности. Например, случайный и независимый выбор объектов, выбор «типичных представителей в каждом классе». Модель извлечения выборки определяет типы возможных ошибок в полученной таблице обучения. От длины выборки зависит качество построенного решающего правила распознавания. Не всегда удаётся получить выборку, имеющую достаточную для получения желаемого качества распознавания длину. Иногда использование длинных выборок может повлечь перенастройку (over fitting) решающих правил. Большие выборки целесообразно разделять на две части: обучающую, по которой происходит индуктивный синтез решающих правил, и контрольную, по которым оценивается качество выбранного решающего правила. Контрольная выборка не участвует в обучении и оценивает единственное решающее правило, найденное на предварительном этапе синтеза.

Метод обучения определяет, как использовать таблицу обучения для выбора экстремального по качеству решающего правила распознавания из некоторого зафиксированного семейства правил. Метод может учитывать целый ряд деталей и использовать различные приёмы. В частности, может учитываться последовательность предъявления объектов таблицы обучения (если от этой последовательности может зависеть результат). Возможно исключение и добавление объектов выборки в процессе обучения. Скользящий контроль также может рассматриваться как метод обучения.

Правило распознавания извлекается в процессе обучения из некоторого заранее зафиксированного семейства правил. Фиксация этого семейства происходит с учётом внешней дополнительной информации о задаче, например, заведомой линейности дискриминантных функций, которые являются геометрическим эквивалентом решающих правил распознавания. Качество извлечённого правила может оцениваться различными способами, например, числом (частотой) ошибок на обучающей выборке, числом ошибок на контрольной выборке, длиной алгоритмического описания найденного решающего правила. Окончательное решение о приемлемости извлеченного правила распознавания принимается на основе анализа всех указанных свойств задачи и задаваемых параметров – требуемой точности, надёжности, алгоритмической сложности извлечённого правила. Указанное выше окончательное решение теоретически может приниматься автоматически, определяя завершение процесса решения задачи распознавания или продолжение поиска с возможными изменениями в выборе фиксируемого семейства решающих правил и других свойств. Однако такой автоматический выбор в настоящее время не разработан на алгоритмическом уровне, поэтому он реализуется исследователями на основе некоторых соображений, сформулированных в процессе решения практических задач распознавания.

Далее предлагается описывать классы задач распознавания в терминах значений свойств, которые представлены в таблице, по схеме:

$$STD/VAR/SFM/SLen/ADI. \tag{1}$$

Запись вида (1) называется кодом задачи распознавания. При невозможности характеризации некоторого свойства задачи в соответствующую позицию кода задачи ставится пропуск. Пропуск означает, что соответствующее свойство может быть любым из его перечисленных значений в таблице, и никакой информации о предпочтительном значении нет. Например, запись

$$D/D_2/R_2T/SS/-$$

определяет детерминированную задачу распознавания с бинарными признаками и случайным, независимым и безошибочным извлечением небольшого числа пар «объект-класс» в таблицу обучения. При этом дополнительная информация для задачи отсутствует. Кроме этого, в синтаксисе кодов будем допускать логические связки «И», «ИЛИ», «НЕ» для комбинированного описания свойств задач. Например, $\bar{L}$ будет обозначать нелинейность; $R_1T \vee R_2T$ – безошибочный выбор объектов из генеральной совокупности с безошибочной их классификацией «учителем» или безошибочный выбор пар «объект-номер класса» из генеральной совокупности пар.

Обозначение свойства	*Наименование свойства*	*Возможные значения свойства задачи: коды и расшифровки*
STD	Стохастичность или детерминированность	S – стохастическая непараметрическая; S_k – стохастическая k-параметрическая; D – детерминированная; ND – недетерминированная.
VAR	Дискретность или непрерывность	D_k – k-значные переменные; C – непрерывные переменные; M – смешанные переменные.
SFM	Модель извлечения и формирования выборки	R_1T – случайный, независимый и безошибочный выбор объектов из генеральной совокупности с безошибочным (при $STD = D$) указанием (учителем) классов, которым эти объекты принадлежат; R_2F – случайный, независимый выбор пар «объект-номер класса» из генеральной совокупности пар с возможными ошибками в любых их компонентах; ST – специальным образом организованное извлечение не содержащей ошибок таблицы обучения (например, выбор типичных объектов или описание прецедентов экспертами); SF – специальным образом организованное извлечение таблицы обучения, возможно с ошибками.
$SLen$	Длина выборки	SS – малая выборка, не допускающая поплнение; AS – выборка средней длины; LS – большая или пополняемая выборка.
ADI	Дополнительная информация о задаче	L – линейность; M — монотонность; CM – компактность классов; RR – наличие областей запрета в признаковом пространстве; SI – другая специальная информация.

Класс задач распознавания $D/-/R_1T \vee R_2T \vee ST/SS/-$

Рассмотрим класс задач распознавания, определяемый кодом

$$D/-/R_1T \vee R_2T \vee ST/SS/-.$$

Значение параметра $SLen = SS$ определяет малую выборку. Малой считается выборка, при обработке которой способами, основанными на статистических методах группировки наблюдений и аппроксимации, невозможно достичь заданной точности

и достоверности. Рассматриваемые задачи – детерминированные с точной обучающей таблицей: каждый объект достоверно принадлежит одному классу (или одновременно нескольким классам, если классы пересекаются). Пересечения классов, не теряя общности, можно считать отдельно выделенными классами.

Выбор методов и моделей решения таких задач определяется следующими соображениями. Разбиение исходной выборки на собственно обучающую и контрольную при решении задач рассматриваемого семейства нецелесообразно по следующим причинам. Точность оценивания, недостижимая на малой выборке, тем более будет недостижимой на ее части; безошибочность информации в выборке и малое число прецедентов делает нецелесообразным отказ от использования всех начальных данных при обучении.

При удачном выборе семейства решающих правил в детерминированных задачах иногда можно указать достаточную длину обучающей выборки для получения точного и единственного решения. В таком случае некорректность по Адамару, свойственная большинству задач распознавания, не будет иметь места. Скользящий контроль нецелесообразен по причине недостаточной представительности начальных данных.

Выбор решающего правила, ошибочно классифицирующего хотя бы один объект таблицы обучения, может повлечь большие ошибки при классификации произвольных допустимых объектов. Поэтому применение корректных алгоритмов и только их приемлемо для решения задач распознавания рассматриваемого класса. Действительно, для рассматриваемых задач некорректность (наличие некоторого числа ошибок на обучающей выборке) влечёт не меньшее число ошибок для извлечённых решающих правил по всей генеральной совокупности. Перечисленные соображения определяют для класса задач

$$D/-/R_1T \vee R_2T \vee ST/SS/-$$

стратегию поиска решающего правила, не допускающего ошибок на объектах таблицы обучения (корректного на выборке) с использованием всей имеющейся выборки при обучении. Поскольку корректные на выборке алгоритмы обучения доставляют нулевое значение эмпирическим функционалам качества обучения, отказ от скользящего контроля становится еще более обоснованным.

При каких же условиях указанная стратегия для класса задач

$$D/-/R_1T \vee R_2T \vee ST/SS/-$$

будет успешной: построенное решающее правило действительно будет обучено распознаванию объектов, не принадлежащих таблице обучения?

Предположим, решающее правило будет выбираться в процессе обучения из семейства правил S . Семейство S обычно определяется методом решения задачи распознавания (методы решения задач распознавания интенсивно разрабатывались в течение прошедшего полувека; например, широко применяются метод вычисления оценок [4], методы обучения нейронных сетей, методы обучения на основе синтеза решающих деревьев и другие [5]). Задачи распознавания, не теряя общности, можно исследовать для случая только двух классов, соответственно, с номерами 0 и 1. Далее будут рассматриваться задачи с двумя классами (единственным основным

свойством). Используя таблицу обучения $\{(X_q, \tilde{\omega}(X_q)), q = 1, \ldots, l\}$, составим функциональную систему

$$\begin{cases} f(X_1) = \alpha_1; \\ f(X_2) = \alpha_2; \\ \ldots\ldots\ldots\ldots \\ f(X_l) = \alpha_l; \\ f \in S, \end{cases} \tag{2}$$

в которой $\alpha_q = \omega(X_q)$ является нулем или единицей – номером класса допустимого объекта X_q. Решением функциональной системы (2), если оно существует, является любое корректное на обучающей выборке решающее правило (функция) $f^* \in S$. Процесс обучения, направленный на поиск корректного на выборке решающего правила, можно рассматривать как поиск решения f^* системы (2). При этом результат, очевидно, определяется выбором семейства S, в рамках которого идет поиск. Выбор корректного решения $f^* \in S$ (решающего правила) будем называть точной настройкой на выборку. Предположим, для любой таблицы обучения с произвольным «столбцом» номеров классов при выбранном семействе S возможна точная настройка, причём существует не единственное корректное (на таблице обучения) правило $f^* \in S$, а обучение происходит по всей выборке и оценивается функционалом эмпирического риска по этой же самой выборке. Тогда никаких гарантий правильного распознавания правилом f^* объектов, не участвовавших в обучении, нет. Действительно, при достаточно «богатом» семействе S можно построить для любого конечного множества, содержащего m допустимых объектов, не участвовавших в обучении, корректную на исходной таблице обучения функцию $f^* \in S$, ошибающуюся на всех этих m объектах. Для этой цели каждому из них сопоставляется неправильный номер класса, и полученная таблица сливается с таблицей обучения $\{(X_q, \omega(X_q)), q = 1, \ldots, l\}$. Если в семействе S найдется корректный алгоритм для такой объединенной таблицы длины $l + m$, то он будет примером случая, когда указанная стратегия обучения в рассматриваемом классе задач, несмотря на точную настройку, даёт неприемлемый результат. В таком случае говорят, что обучаемость не имеет места.

Другая ситуация возникает, когда точная настройка при выбранном семействе S возможна только для некоторого множества допустимых выборок, таких, в которых все объекты в каждом классе обладают некоторыми отличающими их от объектов другого класса свойствами. Тогда возможность получения решения системы (2) с ростом длины выборки l связывается именно с проявлением в выборке указанных свойств (закономерностей – по Колмогорову) и обеспечивается существованием в классе S правила, способного «улавливать» эти свойства. Именно наличие закономерностей в генеральной совокупности в свою очередь влечет появление в таблице обучения не любых из 2^n возможных двоичных «столбцов» $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_q, \ldots, \alpha_l)$, $\alpha_q = \omega(X_q)$, $q = 1, \ldots, l$, а «столбцов» лишь из некоторого, определенного существующей закономерностью, множества.

Далее $VCD(S)$ обозначает емкость семейства решающих правил (размерность Вапника-Червоненкиса [1,2])

Теорема 1. *Если $VCD(S) \geq l$, то найдётся выборка $\{X_1, \ldots, X_l\} = X_l$ такая, что для любого $\tilde{\alpha}$ по таблице $(\tilde{X}_l, \tilde{\alpha})$ возможна точная настройка.*

Доказательство. Следует из определения $VCD(S)$ – ёмкости семейства S □

Теорема 2. *Если для любой выборки $\tilde{X}_l$ существует такой булевский набор $\tilde{\alpha}$, что по таблице$(\tilde{X}_l, \tilde{\alpha})$ невозможна точная настройка, то $VCD(S) < l$.*

Доказательство. Достаточно заметить, что утверждение теоремы 2 равносильно утверждению теоремы 1 □

Последняя теорема дает необходимое условие обучаемости для задач распознавания класса

$$D/-/R_1T \vee R_2T \vee ST/SS/-$$

при выборе стратегии, направленной на построение корректных на обучающих таблицах алгоритмов: емкость семейства решающих правил, используемого для настройки, должна быть меньше длины выборки. Заметим, что неотрицательная величина $l - VCD(S)$ может быть использована для получения оценки неслучайности обнаружения закономерности по обучающей выборке [6]. Условие $VCD(S) < l$ обосновывает важность знания емкости используемого для решения задачи распознавания класса. В связи с этим следующий результат, предполагающий использование хорошо изученных математических свойств некоторых классов функций, представляется полезным при изучении VCD.

Теорема 3. *Пусть функциональная система (2) при зафиксированном семействе решающих функций S для любой выборки $\tilde{X}_l$ и любых двоичных значениях $\alpha_1, \ldots, \alpha_l = \tilde{\alpha}$ может иметь не более одного решения, и при этом найдется выборка $\tilde{X}_l$ такая, что для любого $\tilde{\alpha}$ существует решение f_1. Тогда $VCD(S) = l$.*

Доказательство. Поскольку существует выборка $\{X_1, \ldots, X_l\}$, для которой функциональная система (2) имеет решение при любом двоичном наборе $\alpha_1, \ldots, \alpha_l$, в семействе S найдутся функции, разбивающие эту выборку на два класса всеми способами. Поэтому $VCD(S) \geq l$. Если к любой выборке $\{X_1, \ldots, X_l\}$ длины l добавить один произвольный не принадлежащий ей элемент Z из генеральной совокупности допустимых объектов, то множество решений функциональной системы

$$\begin{cases} f(X_1) = \alpha_1; \\ f(X_2) = \alpha_2; \\ \ldots\ldots\ldots\ldots \\ f(X_l) = \alpha_l; \\ f(Z) = \beta; \\ f \in S, \end{cases} \tag{3}$$

при любом $\beta \in \{0; 1\}$ будет не шире множества решений системы (2). Поэтому, в силу условия теоремы, система (3) может иметь не более одного решения. Если она не имеет решений, то для выборки $\{X_1, \ldots, X_l, Z\}$ длины $l + 1$ при помощи функций семейства S невозможно получить разбиение, соответствующее булеву набору $\alpha_1, \ldots, \alpha_l, \beta$. Если же решение f^* системы (3) существует, то по условию теоремы

оно единственное для функциональных систем (2) и (3). Тогда $\beta = f^*(Z)$, но при помощи функций системы S невозможно получить разбиение выборки $\{X_1, \ldots, X_l, Z\}$ длины $l+1$, соответствующее булеву набору $\alpha_1, \ldots, \alpha_l, \bar{\beta}$. Учитывая, что последнее заключение получено в результате рассмотрения любой выборки $\{X_1, \ldots, X_l\}$ длины l, получаем неравенство $VCD(S) \leq l$, которое вместе с неравенством $VCD(S) \geq l$ дает результат: $VCD(S) = l$ □

Заключение

В статье предложена классификация задач распознавания по их основным свойствам. Обосновывается целесообразность выбора методов решения таких задач в соответствии с особенностями указанных классов. Представлена первая часть исследования в этом направлении и рассмотрен детерминистский класс задач

$$D/-/R_1T \vee R_2T \vee ST/SS/-.$$

Для указанного класса приведено необходимое условие обучаемости и обосновано применение корректных алгоритмов. Дальнейшие исследования в этом направлении связаны с изучением других классов задач распознавания в соответствии с их классификацией.

Список литературы

1. *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979. – 448 с.
2. *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. М.: Наука, 1974. – 416 с.
3. *Воронцов К. В.* Комбинаторные оценки качества обучения по прецедентам // Докл. РАН. – 2004. – Т.394, №2. – С. 175 –178.
4. *Журавлев Ю.И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. – Вып.33. – М.: Наука, 1978. – С. 5 – 68.
5. *Devroye L., Gyorfi L., Lugosi G.* A Probabilistic Theory of Pattern Recognition. Springer-Verlag, NY, 1996. – 636 p.
6. *Donskoy V. I.* The Estimations Based on the Kolmogorov Complexity and Machine Learning from Examples // Proceedings of the Fifth International Conference "Neural Networks and Artificial Intelligence"(ICNNAI'2008). – Minsk: INNS. – 2008. – P. 292 – 297.

Статья поступила в редакцию 20.06.2010