ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.7:007.52;519.81;007.01.362

С.И. Колесникова

МЕТОДЫ АНАЛИЗА ИНФОРМАТИВНОСТИ РАЗНОТИПНЫХ ПРИЗНАКОВ[1]

Рассматривается проблема предварительного анализа информативности признаков, используемых в интеллектуальных системах поддержки принятия решений. Анализируются особенности дискретного и 3-х вероятностных подходов к оценке информативности признаков в условиях их возможной взаимозависимости. Получены двусторонние оценки для распределения минимального числа признаков, обеспечивающих гарантированный уровень качества процедуры распознавания.

Ключевые слова: *интеллектуальная система*, *тестовое распознавание образов*, *весовые коэффициенты признаков*, *энтропия*, *оценки плотности Парзена – Розенблатта.*

Одним из подходов повышения качества распознавания образов и снижения вычислительных затрат является проведение предварительного анализа обучающей информации [1, 2]. Целью такого анализа является оценка основных информативных характеристик обучающей выборки, в частности оценка информативности признаков, оценка значений признаков, выделение наиболее представительных объектов.

С метрической точки зрения различают следующие типы признаков.

Количественные (числовые) признаки – это признаки, замеренные в определенной шкале и в шкалах интервалов и отношений.

Качественные (ранговые, порядковые, балльные) – используются для выражения терминов и понятий, не имеющих цифровых значений (например, тяжесть состояния) и замеряются в шкале порядка.

Номинальные (например, профессия, группа крови, тип хозяйства, национальность, пол) – это признаки, замеренные в шкале наименований. При анализе таких признаков каждую отметку номинальной шкалы считают отдельным самостоятельным признаком. Он принимает всего два значения (как правило, используют значения 0 и 1), разность которых можно интерпретировать как степень важности несовпадения данного признака при сравнении двух объектов. Такие признаки называют двоичными, бинарными, дихотомическими признаками [4, 5 – 9]. После проведения дихотомизации [5] (преобразования исходных показателей в набор признаков с двумя градациями) номинальные измерения становятся доступны для

[1] Работа выполнена при финансовой поддержке РФФИ (проекты № 09-01-99014-р_офи, № 07-01-00452).

применения широкого спектра различных методов многомерного количественного анализа с учетом специфики данного вида измерений.

В практических задачах распознавания первоначальные описания объектов содержат все доступные наблюдению (или измерению, вычислению) характеристики или параметры, поэтому в описании объектов участвуют несколько десятков (сотен) величин (например, в задачах медицинской диагностики, геологического, технического и социологического прогнозирования и т.д.). Однако набор большого числа прецедентов часто требует дорогостоящих и трудоемких работ (в некоторых случаях медицинской диагностики, прогнозирования редких металлов вообще невозможен) [1 – 3]. В прикладных работах наиболее распространенный пример объектов нечисловой природы – разнотипные данные. Так, в реляционной базе данных (как разновидности обучающего множества) информация о реальном объекте описывается вектором, часть координат которого – значения количественных признаков, а часть – качественных (номинальных и порядковых).

При принятии решения о выборе класса, которому принадлежит анализируемый объект, возникает проблема его оценки по нескольким признакам и корректного учета этих оценок при их обобщении или совместном рассмотрении на этапе принятия итогового решения [1 – 7]. Эта проблема в свою очередь подразделяется на две подпроблемы: установление вида решающего правила (так называемого «обобщенного признака») и определение весовых коэффициентов – коэффициентов важности «частных» (характеристических) признаков [1–7, 9], отражающих свойства объектов.

В данной работе рассматриваются особенности дискретного подхода к оценке информативности разнотипных признаков и анализируются статистические подходы, предложенные в [12, 16]. Применимость вероятностных и статистических подходов к оцениванию информативности признаков в интеллектуальных системах принятия решений возможна в случае пропусков данных; большой размерности задачи и пр. и никак не связана с тем, в какой шкале измеряются значения оцениваемых признаков.

1. Основные определения и понятия

В общих чертах задача распознавания состоит в следующем [1 – 3, 7, 9].

Исследуется некоторое множество объектов $O = \{O_1, ..., O_m\}$. Известно, что O представимо в виде объединения K подмножеств $S_1, \ldots, S_K$, называемых классами. Наиболее общим определением класса является следующее [7]: класс – это совокупность (семейство) объектов, обладающих общими свойствами. Информация о свойствах объекта может быть получена путем наблюдений, измерений, оценок и т.п. и представлена совокупностью признаков, значения которых выражаются в числовых и/или вербальных шкалах. Входящие в один тот же класс объекты считаются неразличимыми (эквивалентными), а каждый класс объектов характеризуется некоторым качеством, отличающим его от других классов. Вместе все классы должны составлять исходную совокупность объектов.

Наряду с классом используется понятие образа [9]. В матричной модели представления данных и знаний, включающей матрицу описаний (Q) объектов в пространстве характеристических признаков и матрицу различений (R) объектов в пространстве классификационных признаков [9], множество всех неповторяющихся строк матрицы R сопоставлено множеству выделенных образов. Элементами образа являются объекты, представленные строками матрицы Q, сопоставленными одинаковым строкам матрицы R.

Если имеется единственный механизм классификации, матрица различений вырождается в столбец, что соответствует традиционному представлению знаний в задачах распознавания образов [1, 9].

Объекты из O описываются некоторой системой признаков $\{z_1, \ldots, z_M\}$. Имеется конечный набор $O_1,\ldots, O_m$ объектов из O, о которых известно, каким классам они принадлежат. Это прецеденты, или обучающие объекты. Пусть их описания имеют вид $q_1 = (q_{11}, \ldots, q_{1M})$, $q_2 = (q_{21}, \ldots, q_{2M})$, …, $q_m = (q_{m1}, \ldots, q_{mM})$, здесь q_{ij} – значение признака z_j для объекта O_i. Требуется по предъявленному набору значений признаков $(b_1, \ldots, b_M)$, описывающему некоторый объект из O (не входящий в обучающее множество $O_1,\ldots, O_m$), о котором неизвестно, какому классу он принадлежит, определить этот класс.

2. Постановка задачи

Требования к подготовке данных (в соответствии с целями исследования) для математического анализа, для оценки характера распределения в исследуемой выборке (предварительный анализ данных) диктуют выполнение следующих работ [1 – 7]:

1) проверку однородности выбранных групп наблюдения, в том числе контрольных, что может быть проведено или экспертным путем, или методами многомерной статистики (например, с помощью кластерного анализа);

2) нормализацию переменных, т.е. устранение аномалий показателей в матрице данных (согласование мнений);

3) снижение размерности пространства признаков (формальными методами путем оценки информативности);

4) стандартизованное описание признаков;

5) построение классификационных шкал признаков, т.е. процедуру идентификации и установления физических границ изучаемых параметров и представление информации в квантованной форме (каждому значению признака соответствует определенное кодовое число).

Поставим, во-первых, задачу определения формальных подходов математической обработки и оценки информативности разнотипных признаков (п.п. 1, 3) – наиболее важных из вышеперечисленных как на этапе предварительного (для оценки характера распределения в исследуемой выборке), так и окончательного анализа в соответствии с целями исследования; во-вторых, задачу оценивания вероятностного распределения минимально необходимого числа информативных признаков, обеспечивающих заданный уровень надежности алгоритма (процедуры) распознавания, если известно вероятностное распределение признаков (что естественно для приложений, где накоплены соответствующие статистические данные).

3. Методы оценки информативности разнотипных признаков

Рассматриваются пять подходов к отбору информативных признаков: на основе дискретных методов поиска в обучающей выборке информативной зоны [2, 3]; на основе методов кластеризации [5 – 7, 10]; на основе предположения о нормальности распределений объектов в кластерах [11 – 14]; на основе теоретико-информационного понятия энтропии [15, 16]; на основе непараметрических оценок плотности [19].

3.1. Нахождение в обучающей выборке информативной зоны

В [2, 3] описывается методика предварительного анализа обучающей информации, основанная на нахождении в обучающем материале информативной зоны (информативными считаются такие подописания (или фрагменты описаний), которые позволяют различать объекты из разных классов или отличать данный объект от всех объектов, не принадлежащих тому же классу, что и рассматриваемый) и типичных для своих классов объектов (наиболее представительных). Информативная зона выделяется на основе оценки типичности значений каждого признака. Предлагается два способа выделения типичных объектов: 1) типичными считаются объекты, описания которых состоят из типичных значений признаков; 2) типичными являются объекты, которые правильно распознаются на скользящем контроле. Так, к нетипичным («шумящим») признакам относятся признаки, принимающие много значений или значения которых редко встречаются во всех классах (про такие признаки нельзя сказать, что они являются значимыми). Каждый объект обучающей выборки, лежащий на границе между классами, также не является «типичным» для своего класса, поскольку его описание похоже на описания объектов из других классов.

Поиск информативных фрагментов основан на использовании аппарата дискретной математики, в частности булевой алгебры, теории дизъюнктивных нормальных форм, теории покрытий булевых и целочисленных матриц. Решению этой задачи посвящены работы [1 – 3, 9]. Так, в алгоритмах вычисления оценок, разработанных Журавлевым Ю.И. и его учениками [1 – 3], находятся оценки ансамблей признаков, которые являются обобщениями коэффициентов информативности, рассмотренных в [2].

Для тестовых алгоритмов распознавания образов в [2] введена мера важности признака (его информационный вес) как отношение числа вхождений признака во все безызбыточные (тупиковые) тесты к числу всех безызбыточных тестов. В работе [9] на основе матричной модели представления данных и знаний получена формула, выражающая числовую оценку различающей способности признака (весовой коэффициент признака).

Перечисленные в [1 – 3, 9] методы оценки информативности признаков показывают хорошие результаты в прикладных задачах. При этом под качеством распознавания понимается качество алгоритма вне обучающей выборки (способность алгоритма к обобщению или экстраполяции), которое оценивается долей (в процентах) правильно распознанных объектов при проведении процедуры скользящего контроля [10].

Отметим, что применение дискретного подхода оказывается во многих случаях сложным в силу чисто вычислительных трудностей переборного характера, возникающих на этапе поиска информативных фрагментов описаний объектов [3]. При числе признаков, равном M, число непустых подмножеств составляет 2^{M-1} и прямой перебор всех подмножеств оказывается невозможным уже при M порядка 20 даже на самых современных компьютерах. В силу экспоненциального роста числа фрагментов при возрастании размерности описаний, решение проблемы только за счет повышения производительности вычислительной техники нереально, поэтому дискретные методы трудно применимы для предварительного анализа большого объема обучающей информации.

Наиболее полное изложение методов дискретного анализа информации в задачах распознавания можно найти в [1 – 3, 10].

3.2. Отбор информативных признаков на основе методов кластеризации

Кластеризационные методы отбора признаков позволяют разбить выборку признаков на кластеры, состоящие из схожих признаков, и выделить в каждой группе по одному наиболее типичному представителю. Для применения методов кластеризации, рассмотренных, например, в [4 – 6, 10, 11], необходимо ввести метрику на множестве признаков Z. Рассмотрим возможные метрики на признаках.

Пусть z_1, z_2 – два произвольных признака из Z. Выборочные векторы значений признаков (для объектов выборки O объема n) обозначим соответственно через $z_1 = (z_1(O_1), z_1(O_2), ..., z_1(O_n))$ и $z_2 = (z_2(O_1), z_2(O_2), ..., z_2(O_n))$. Рассмотрим три варианта определения функции расстояния $\rho(z_1, z_2)$.

1. *Метрика на основе коэффициента линейной корреляции* $r(z_1, z_2)$, применимая для количественных признаков:

$$\rho(z_1, z_2) = 1 - |r(z_1, z_2)|, \; r(z_1, z_2) = \sum_{i=1}^{n} z_1'(O_i) \cdot z_2'(O_i),$$

где $z_1'(O_i)$, $z_2'(O_i)$ – нормированные и центрированные значения признаков z_1, z_2. Расстояние $\rho(z_1, z_2) = 0$ тогда и только тогда, когда признаки связаны линейной зависимостью.

2. *Метрика Кендалла – Кемени*, применимая для порядковых признаков. Она определяется как доля пар объектов O_i, O_j с различными порядковыми отношениями между значениями признаков z_1 и z_2:

$$\rho(z_1, z_2) = \frac{1}{2 \cdot C_n^2} \sum_{i=1}^{n-1} \sum_{j=i}^{n} \left| \text{sign}(z_1(O_i) - z_1(O_j)) - \text{sign}(z_2(O_i) - z_2(O_j)) \right|.$$

Это расстояние равно нулю тогда и только тогда, когда признаки связаны монотонной зависимостью, то есть существует монотонная функция f, такая, что $z_1(O_j) = f\left(z_2(O_j)\right)$ ($j = 1, \ldots, n$).

3. *Метрика Хэмминга*, применимая для номинальных признаков с одинаковыми множествами допустимых значений $D_{z_1} = D_{z_2}$, которая обращается в нуль тогда и только тогда, когда векторы z_1 и z_2 совпадают:

$$\rho(z_1, z_2) = \sum_{i=1}^{n} [z_1(O_i) \neq z_2(O_i)].$$

Для номинальных признаков с различными множествами значений ищется соответствие σ: $D_{z_1} \to D_{z_2}$, при котором расстояние Хэмминга минимально (без ограничения общности предполагается, что $|D_{z_1}| \geq |D_{z_2}|$):

Если необходимо найти метрику между разнотипными признаками, измеренными в разных шкалах, то они сначала приводятся к одной общей шкале [6 – 8, 11].

Отметим следующие особенности применения методов кластеризации.

Во-первых, если в исходном множестве признаков будут присутствовать неинформативные признаки, то могут появиться кластеры, целиком состоящие из неинформативных признаков.

Во-вторых, вышеупомянутые метрики используют информацию о попарном сходстве между признаками, но не решают проблему мультиколлинеарности признаков (набор попарно некоррелированных признаков может оказаться линейно зависимым).

В-третьих, классификация по принципу минимального расстояния пригодна, если кластеры, соответствующие разным классам, разнесены достаточно далеко друг от друга [7], т.е. значения признаков, описывающих объекты из разных образов, достаточно далеки друг от друга.

3.3. Статистический кластер-алгоритм

Алгоритм построения кластеров в предположении о нормальности распределений объектов в кластерах (что является естественным во многих практических приложениях) использует кластер-критерий, предложенный в [12], и свойство оценок Мешалкина Л.Д. [13] оценивать параметры распределения кластера, наибольшего по количеству точек. Отметим, что условия критерия К. Пирсона [14] проверки гипотезы о совпадении центров двух многомерных нормальных совокупностей с известными ковариациями, различающихся сдвигом, являются непригодными для кластер-анализа в силу жесткости используемых предположений.

Идея построения кластер-алгоритма, разбивающего выборку признаков на кластеры, состоящие из схожих признаков, заключается в следующем.

Кластер-критерием на уровне значимости α проверяется гипотеза о принадлежности выборки одному нормальному распределению. Если гипотеза принимается, то выборка с вероятностью $1 - \alpha$ содержит один кластер. Если гипотеза отвергается, то самый большой кластер вырезается эллипсоидом на уровне значимости α и к оставшимся точкам применяется описанная процедура в цикле. Процедура выделения кластеров заканчивается, когда все точки распределены по кластерам либо их количество невелико (α-остатки от проверки гипотез).

В отличие от большинства алгоритмов кластерного анализа данный статистический кластер-алгоритм не требует предварительного указания числа K классов, на которое надо поделить совокупность точек наблюдения, и решает задачи со сложными наложениями и пересечениями нескольких классов (случай, когда признаки могут принимать одинаковые значения для объектов из разных образов).

3.4. Методы оценки информативности признаков на основе энтропии

В терминах теории информации мерой трудности распознавания служит энтропия H распределений плотности вероятности образов [6, 15, 16].

1. *Метод на основе формализма мультимножеств и теоретико-информационном понятии энтропии.* Основанный на формализме мультимножеств [17] метод определения весовых коэффициентов признаков при принятии решения в интеллектуальных тестовых распознающих системах с матричным представлением данных и знаний [9] учитывает вклад признаков в распознающую способность теста с учетом их взаимозависимости и базируется на представлении совокупности всех различимых пар объектов из разных образов для каждого признака z_i, $i = \overline{1, M}$, в виде мультимножества [18].

Поставим данному признаку z_i в соответствие совокупность различимых i-м признаком пар объектов из разных образов и будем говорить, что признак z_i по-

рождает мультимножество $\{k_{P_i}(u) \bullet u \mid u \in U,\ k_{P_i}(u) \in Z_+\}$. Таким образом, представленные признаки [17] являются множествами с повторяющимися элементами $u \in U$ (мультимножествами), при этом мощность мультимножества определяется как общее число его элементов, где множество U – домен или универсальное множество [17], откуда «черпают» свои возможные значения мультимножества, порожденные соответствующими признаками. Заметим, что элементы домена u являются обозначениями пар объектов из разных образов, поэтому элементы (i–j) и (j–i) (пары номеров объектов) считаются эквивалентными ($i \neq j$, $i, j \in \{1, \ldots, K\}$).

Обозначим через $P(v/i)$ вероятность «проявления» v-го элемента домена для i-го образа, тогда вероятность v-го элемента домена для всех K образов равна $P_v = \sum_{i=1}^{K} P(v/i)$. С учетом условия нормировки доля i-го образа в этой сумме $r_i = P(v/i)/P(v)$ и энтропия v-го элемента домена выражается следующим значением: $H_v = -\sum_{i=1}^{K} r_i \cdot \log_2 r_i$.

Из свойства аддитивности энтропии следует, что общая неопределенность при распознавании образов по признаку z_i имеет вид $H_{z_i} = -\sum_{v=1}^{N_U} H_v \cdot P_v$. Количество информации, получаемой в результате измерения признака z_i, равно $H_0 - H_{z_i}$, H_0 – исходная неопределенность относительно образов $H_0 = \log_2 K$.

При качественном анализе эмпирических данных роль вероятностей $P(v/i)$ (вероятность «проявления» v-го элемента домена для i-го образа) играют их выборочные оценки (частости, доли).

В реальных таблицах данных зависимость между признаками наблюдается очень часто. А если признаки зависимы, то при выборе наиболее информативной подсистемы признаков оценками их индивидуальной информативности руководствоваться нецелесообразно.

2. *Анализ взаимосвязи признаков на основе энтропии*. Трактовка статистической связи между переменными z_i и z_j сводится к оценке количества информации $I(z_i, z_j) = H(z_j) - H(z_j / z_i)$, которое уменьшает неопределенность того, какое значение примет z_j, если известно значение z_i. В качестве оценки меры сопряженности примем долю сопряженного разнообразия величины z_j, включенной в систему (z_i, z_j), по сравнению с разнообразием (энтропией) z_j, рассматриваемой отдельно: $I(z_i, z_j)/H(z_j)$, где в явной форме величина $I(z_i, z_j)$ определяется по формуле

$$I(z_i,\ z_j) = -\sum_{s=1}^{l}\sum_{k=1}^{l} p(x_s, y_k) \log_2 \frac{p(x_s \mid y_k)}{p(x_s)} = -\sum_{s=1}^{l}\sum_{k=1}^{l} p(x_s, y_k) \log_2 \frac{p(x_s, y_k)}{p(x_s) \cdot p(y_k)},$$

где l – число градаций признаков.

Отметим, что в силу симметричности выражения $I(z_i, z_j)$ относительно z_i, z_j значения количества информации, заключенного в признаке z_i о признаке z_j и заключенного в признаке z_j о признаке z_i, равны [14], т.е. среднее количество информации есть мера соответствия двух признаков, характеристика их связи, а не характеристика одного из признаков.

Взаимная информация $I(z_i, z_j)$ обращается в нуль тогда и только тогда, когда признаки z_i и z_j статистически независимы. Максимальное значение взаимной информации, равное $H(z_i)$ или $H(z_j)$, соответствует функциональной зависимости (полной связи) признаков z_i и z_j, когда каждому значению признака z_i соответствует единственное значение признака z_j. Заметим, что использование коэффициента $I(z_i, z_j)/H(z_j)$ для разделения признаков на «зависимые» и «определяющие», вообще говоря, не имеет достаточного обоснования.

Отметим, что построение теоретико-информационной меры связи может осуществляться не только на основе энтропии. Важным примером такой меры служит коэффициент Валлиса, реализующий принцип «пропорциональной предикции», согласно которому мерой связи должно служить относительное уменьшение вероятности ошибки предсказания признака z_j при знании признака z_i в сравнении с вероятностью ошибки прогноза z_j без знания z_i. Интерпретация коэффициента Валлиса весьма проста: если, например, его значение равно 0,5, то знание z_i уменьшает число ошибок прогноза значения z_j вдвое. Однако в выборочных исследованиях предпочтительнее пользоваться не коэффициентом Валлиса, а информационными мерами связи, значимость которых может быть установлена, например, в соответствии с критерием Пирсона [14].

3.5. Оценка информативности разнотипных признаков на основе непараметрических оценок плотности

Методы оценивания плотности вероятности в пространствах общего вида предложены и первоначально изучены в [19]. В частности, в задачах классификации объектов нечисловой природы предлагается использовать непараметрические ядерные оценки плотности типа Парзена – Розенблатта (этот вид оценок и его название введены в [19] по имени американских ученых Парзена и Розенблатта, ранее использовавших подобные статистики в случае $X = R^1$, $\rho(x_i, x) = |x_i - x|$):

$$f_n(x) = \frac{1}{v_n(h_n, x)} \sum_{1 \le i \le n} K\left(\frac{\rho(x_i, x)}{h_n}\right),$$

где K: $R_+^1 \to R^1$ – ядерная функция; $\{x_1, x_2, ..., x_n\} \in X$ – выборка, по которой оценивается плотность; $\rho(x_i, x)$ – расстояние между элементом выборки x_i и точкой x, в которой оценивается плотность; последовательность показателей размытости $\{h_n\}$ такова, что при $n \to \infty$ $h_n \to 0$ и $nh_n \to \infty$; $v_n(h_n, x)$ – нормирующий множитель, обеспечивающий выполнение условия $\int_X f_n(x) d\mu = 1$.

В [19] показано, что оценка плотности типа Парзена – Розенблатта $f_n(x)$ является состоятельной ($f_n(x) \to f(x)$ по вероятности при $n \to \infty$) и оценена среднеквадратическая скорость сходимости ядерных оценок $\alpha_n = \mathrm{E}(f_n(x) \to f(x))^2$.

Поскольку пространство разнотипных признаков – это декартово произведение непрерывных и дискретных пространств, то для случая фиксированного числа градаций качественных признаков непараметрическую оценку плотности можно свести к произведению частоты попадания в точку в пространстве качественных признаков на классическую оценку Парзена – Розенблатта [19] в пространстве количественных признаков. Расстояние $\rho(x, y)$ можно рассматривать как сумму

евклидова расстояния ρ_1 между количественными факторами, расстояния ρ_2 между номинальными признаками ($\rho_2(x,y)=0$, если $x \neq y$, и $\rho_2(x,y)=1$, если $x=y$), и расстояния ρ_3 между порядковыми переменными (если x и y – номера градаций, то $\rho_3(x,y)=|x-y|$).

4. Двусторонние оценки для вероятности минимального числа признаков, обеспечивающих гарантированный уровень качества процедуры распознавания

Рассмотрим подход к оценке качества алгоритма распознавания образов как функции «качества» обучающей выборки (объема и содержания выборки в виде «объект-признак»). Основу данного подхода к оценке качества процедуры распознавания составляет заданная информация о частотах встречаемости признаков (оценках вероятностей проявления признака) для каждого из классов (образов). Получим двусторонние оценки для вероятности минимального числа признаков, обеспечивающих гарантированный уровень качества процедуры распознавания, при этом будем использовать вероятностную модель пересечения случайным процессом определенного уровня.

Пусть $\xi_i = \xi_i(m)$ – случайная величина (с.в.), означающая число корректно распознанных объектов на обучающей выборке признаком z_i, $i=\overline{1,n}$ (n – число признаков, m – число объектов).

Случайная величина $\Phi_n = \sum_{i=1}^{n} \xi_i \Big/ nm$ представляет собой долю корректно распознанных объектов по n первым признакам, ранжированным каким-либо образом, например, по убыванию весовых коэффициентов признаков, вычисленных по обучающей выборке (например, по формулам, предложенным в [20]), и является числовой интерпретацией меры надежности распознавания, или качества работы алгоритма распознавания [10, 11].

Введем с.в. (момент остановки случайного процесса Φ_n [21, 22]):

$$\tau = \inf\{n \geq 1 : \Phi_n > q_0\},$$

где q_0 – заданный уровень надежности (качества) процедуры (алгоритма) распознавания.

Поставим задачу оценить вероятностное распределение $R_n = P(\tau > n)$ величины τ, или минимально необходимого числа информативных признаков, обеспечивающего заданный уровень надежности распознавания, если известно вероятностное распределение признаков.

Знание распределения величины τ, являющейся пороговым значением качества алгоритма распознавания, позволит: а) оценить требуемое количество признаков (n) для достижения заданного уровня качества распознавания; б) оптимизировать процедуру распознавания по параметрам n, m, а также их весовым коэффициентам. Так, в приложениях (геологических, медицинских и пр.) актуально решение вопроса о соотношении длины выборки данных (m) (например, дорогостоящее бурение дополнительной скважины) и числа характеристических признаков (n) (введение менее дорогостоящего дополнительного анализа для уже имеющихся данных).

Обозначим

$$\eta_i(k) = \mathrm{P}(\xi_i = k) = \sum_{(j_1,\ldots,j_k)\in J_k} \prod_{l=1}^{k} p_{j_l i} \prod_{t\in\{1,\ldots,m\}\setminus J_k} (1-p_{ti}),$$

$$\nu_i = \mathrm{E}\xi_i = \sum_{k=1}^{m} k\cdot\eta_i(k), \quad \sigma_i^2 = \mathrm{D}\xi_i = \sum_{k=0}^{m} (k-\nu_i)^2\cdot\eta_i(k)$$

и сформулируем результат проведенного исследования.

Теорема. Пусть признак z_i, $i=\overline{1,n}$, распознает объект $O_j = \{a_1^j, \ldots, a_n^j\}$, $j=\overline{1,m}$, с вероятностью p_{ji} и справедливо условие

$$\kappa \mid q_0 m - \nu_i \mid < \varepsilon_c \sigma_i^2 L_i^{-1},\ \varepsilon_c = \ln c \cdot c^{-1}/2,\ i=\overline{1,n},$$

где κ – некоторая положительная величина, 0<κ<1. Тогда имеют место оценки для надежности распознавания $R_n = P(\tau > n)$ (вероятности обеспечения заданного качества распознавания):

$$R_n \le c_1 \exp\left\{-c_2 (q_0 m)^{-2}\left[\sum_{j=1}^{n}\sigma_j^2 c^{-1/2}\right] + c_3\sum_{i=1}^{n}\upsilon_i\right\},$$

$$R_n \ge \exp\left\{-c_4\left[\alpha^2 + (q_0 m)^{-2} c\sum_{j=1}^{n}\sigma_j^2\right] - c_5\sum_{i=1}^{n}\upsilon_i\right\},$$

где $\upsilon_i = (q_0 m - \nu_i)/\sigma_i^2$, величины c_i, $i=\overline{1,5}$, не зависят от n и являются известными функциями от величин (α, p, q, c):

$$c_1 = (5/2\alpha)^{1/2p},\ c_2 = \pi^2(1-3\alpha^{1/2})c^{-1/2}/(8p),\ c_3 = (q-1)c^2,$$

$$c_4 = \pi^2 p(1+15\alpha^{3/4})/8,\ c_5 = c^2\frac{q+1}{q-1},$$

$$\alpha \le 1/16,\ \ 1/p + 1/q = 1,\ \ p>1,\ \ c>1.$$

Заключение

Выбор полезной информации (отбор признаков) является важной операцией, поскольку для решения любой классификационной задачи целесообразно отобрать признаки, несущие не «шум» и не иррелевантную (не относящуюся к цели исследования), а полезную для данной задачи информацию. Для трех статистических подходов оценки информативности признаков, рассмотренных в статье, произведен сравнительный анализ (по числу ошибок классификации) на данных известного репозитория [24], подтвердивший экспериментально целесообразность использования в решающих правилах не все исходные признаки, а наиболее информативные

Если имеется несколько различных методов обучения (соответственно несколько алгоритмов), то в задаче выбора модели (в отсутствии других априорных предпочтений) можно выбирать ту модель, которая обладает лучшей обобщающей способностью [10] (лучшей результативностью на независимой выборке). Дело в том, что по мере увеличения числа используемых признаков (сложности

модели) средняя ошибка на обучающей выборке, как правило, монотонно убывает. При этом средняя ошибка на независимых контрольных данных сначала уменьшается, затем проходит через точку минимума и далее только возрастает. Это явление называют переобучением [10, 11, 23]. Однако получение теоретических оценок вероятности переобучения (отклонение частот ошибок алгоритма на контрольной и обучающей выборке) пока остаётся открытой проблемой [23].

За оценку информативности признака в распознавании образов принято считать отношение качества распознавания контрольной выборки в полном пространстве признаков к качеству распознавания, проводимого без учета оцениваемого признака [1 – 3, 10, 11] (метод скользящего контроля). Если известен вид функции распределения, информативность признаков может оцениваться через средние квадратичные отклонения [5].

Выбор метода исследования на реальных базах данных и знаний целесообразно осуществлять пропорционально эффективности соответствующего метода на контрольной выборке.

ЛИТЕРАТУРА

1. *Журавлев Ю.И.*, *Гуревич И.Б.* Распознавание образов и анализ изображений // Искусственный интеллект: В 3 кн. Кн 2. Модели и методы: Справочник / Под ред. Д.А. Поспелова. М.: Радио и связь, 1990. С. 149–190.
2. *Дмитриев А.И.*, *Журавлев Ю.И.*, *Кренделев Ф.П.* О математических принципах классификации предметов или явлений // Дискретный анализ. Новосибирск: ИМ СО АН СССР, 1966. Вып. 7. С. 1 – 17.
3. *Дюкова Е.В.*, *Песков Н.В.* Построение распознающих процедур на базе элементарных классификаторов. URL: www.ccas.ru/frc/papers/djukova05construction.pdf.
4. *Айзерман М.А.*, *Алескеров Ф.Т.* Выбор вариантов: основы теории. М.: Наука, 1990. 136 с.
5. *Миркин Б.Г.* Анализ качественных признаков и структур. М.: Статистика, 1980. 317 с.
6. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики, 1999. 270 с.
7. *Ту Дж.*, *Гонсалес Р.* Принципы распознавания образов. М.: Мир, 1978. 415 с.
8. *Лбов Г.С.* Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981.
9. *Yankovskaya A.E.* Test pattern recognition with the use of genetic algorithms // Pattern Recognition and Image Analysis. 1999. V. 9. No. 1. P. 121 – 123.
10. *Воронцов К.*В. Обзор современных исследований по проблеме качества обучения алгоритмов // Таврический вестник информатики и математики. 2004. № 1. С. 5 – 24. URL: http://www.ccas.ru/frc/papers/voron 04twim.pdf.
11. *Воронцов К.В.* Лекции по методам оценивания и выбора моделей. 2007. URL: www.ccas.ru/voron/download/Modeling.pdf.
12. *Шурыгин А.М.* Статистический кластер-алгоритм // Математические методы распознавания образов: Сб. докл. 13-й Всерос. конф. Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007 г. М.: МАКС Пресс, 2007. С. 241 – 242.
13. *Meshalkin L.D.* Some mathematical methods for the study of noncommunicable diseases // Proc. 6-th Intern. Meeting of Uses of Epidemiol. in Planning Health Services. Yugoslavia, Primosten, 1971. V. 1. P. 250 – 256.
14. *Pearson K.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling // Phil. Mag. 1900. V. 50. P. 157 – 175.
15. *Перегудов Ф.И.*, *Тарасенко Ф.П.* Введение в системный анализ. М.: Высшая школа, 2001. 396 с.
16. *Колесникова С.И.*, *Янковская А.Е.* Статистический подход к оцениванию зависимых признаков в интеллектуальных системах // Математические методы распознавания об-

разов: Сб. докл. 13-й Всерос. конф. Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007 г. М.: МАКС Пресс, 2007. С. 143 – 146.
17. *Петровский А.Б.* Упорядочивание и классификация объектов с противоречивыми признаками // Новости искусственного интеллекта. 2003. № 4. С. 34 – 43.
18. *Янковская А.Е.*, *Колесникова С.И.* О применении мультимножеств к задаче вычисления весовых коэффициентов признаков в интеллектуальных распознающих системах // Искусственный интеллект. Украина, Донецк: «Наука і освіта», 2004. № 2. С. 216 – 220.
19. *Орлов А.И.* Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 58 – 92.
20. *Колесникова С.И.*, *Янковская А.Е.* Оценка значимости признаков для тестов в интеллектуальных системах // Изв. РАН. Теория и системы управления. 2008. № 6. С. 135 – 148.
21. *Новиков А.А.* О времени выхода сумм ограниченных случайных величин из криволинейной полосы // Теория вероятностей и ее применения. 1981. Т. 26. № 2. С. 287 – 301.
22. *Ширяев А.Н.* Вероятность. М.: Наука, 1980. 575 с.
23. *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. 2008. V. 18. No. 2. P. 243 – 259.
24. *Merz C.J.*, *Murphy P.M.* UCI Repository of machine learning datasets // Information and Computer Science University of California, Irvine, CA, 1998. URL: http://www.ics.uci.edu/~mlearn/databases.

Статья представлена кафедрой программирования факультета прикладной математики и кибернетики Томского государственного университета. Поступила в редакцию 9 октября 2008 г.