Устойчивость и обобщенные оценки классифицированных объектов в разнотипном признаковом пространстве

Н. А. Игнатьев, Ш. Ф. Мадрахимов
Национальный университет Узбекистана, Ташкент
e-mail: n_ignatev@rambler.ru, mshavkat@yandex.ru

Рассматривается вычисление структурных характеристик классифицированных объектов в разнотипном признаковом пространстве. Для вычисления используются методы интеллектуального анализа данных.

Ключевые слова: искусственный интеллект, обобщенные оценки, устойчивость, интервалы доминирования, интеллектуальный анализ данных.

Введение

Разнотипность признаков в описании объектов не позволяет применять в качестве инструментария для исследования методы статистического разведочного анализа данных. Для решения этой проблемы предлагается использовать методы интеллектуального анализа данных, ориентированные на поиск скрытых закономерностей по базам данных.

Одним из направлений интеллектуального анализа является классификация. Значительный объем информации при решении задач классификации представляет знание о структурном размещении объектов классов и сложности конфигурации границ классов.

Сведения о структурном размещении объектов классов в признаковом пространстве по заданной метрике пытались получить различными способами. Например, о сложности конфигурации границ классов можно было судить по результатам корректного распознавания объектов с помощью линейных и кусочно-линейных решающих функций [1]. Другой используемой структурной характеристикой была устойчивость объектов в непересекающихся классах. Задача вычисления устойчивости как меры структурного разнообразия рассматривалась в рамках непараметрических методов распознавания [2].

Устойчивость отображает локальные свойства объектов классифицированной выборки. Знание этих свойств необходимо для определения аномальных объектов классов, объяснения причин выбора объектов в состав эталонов минимального покрытия обучающей выборки, достаточного для ее корректного распознавания.

Многообразие значений устойчивости объектов классов в [2] напрямую зависело от выбора метрики. Поскольку в разнотипном признаковом пространстве отсутствуют меры близости со свойствами метрики, возникла необходимость в использовании иных подходов. Так, характеристика структурного размещения каждого из объектов-эталонов локально-оптимального покрытия $\Pi_j = \{S^1, ..., S^p\}$, $p > 1$, классов обучающей выборки в искусственных нейронных сетях (ИНС) с минимальной конфигурацией [3]

вычислялась через долю некорректно распознанных объектов при скользящем экзамене на множестве Π_j. Решение задачи оценки устойчивости и алгоритмического (без участия экспертов) ранжирования объектов классов по обобщенным оценкам в разнотипном признаковом пространстве ранее не рассматривалось.

В настоящей работе предлагается метод вычисления непересекающихся интервалов количественных признаков, в границах которых доминируют значения тех или иных классов. На базе этого метода стало возможным вычисление как обобщенных оценок объектов в разнотипном признаковом пространстве, так и меры их устойчивости. В моделях ИНС значение меры устойчивости представляет локальный индикатор способности нейронов сети к обобщению в той или иной области разнотипного признакового пространства. Под обобщением понимается корректное (без ошибок) распознавание объектов, которых ИНС "не видела" в процессе обучения.

1. Выбор интервалов доминирования значений количественных признаков классов

Рассматривается множество M допустимых объектов, разбитое на l непересекающихся подмножеств (классов) $K_1, K_2, ..., K_l$, $M = \bigcup_{i=1}^{l} K_i$. Считается, что представители классов заданы через выборку (подмножество M) объектов $E_0 = \{S_1, ..., S_m\}$. Объекты выборки описываются с помощью n разнотипных признаков, множества допустимых значений ξ из которых измеряются в интервальных шкалах, $(n - \xi)$ — в номинальной шкале.

Вычисление обобщенных оценок и устойчивости объектов производится относительно отдельных классов. Необходимость сведения решения к двухклассовой задаче распознавания с объектами из K_t и $CK_t = M \backslash K_t$, $t = \overline{1,l}$, связана со следующими факторами:

— любая обобщенная оценка (показатель) относительна. Объекты каждого из классов противопоставляются объектам противоположных классов (например, класс заболевших и умерших от гриппа и класс практически здоровых людей);

— отсутствуют классы аналитических функций для восстановления зависимостей в пространстве разнотипных признаков.

Требуется:

— на множестве допустимых значений каждого из количественных признаков провести разбиение на минимальное число непересекающихся интервалов с доминированием значений из классов K_t или CK_t, $t = \overline{1,l}$;

— дать количественную оценку структурного размещения каждого из объектов E_0 относительно класса K_t, $t = \overline{1,l}$.

Обозначим через I, J множество номеров соответственно количественных и номинальных (качественных) признаков $X = \{x_1, ..., x_n\}$ в описании допустимых объектов, $\mid I \mid + \mid J \mid = n$. Произведем выбор интервалов для каждого количественного признака, в границах которых доминируют значения классов K_t или CK_t. Для этого упорядочим значения c-го признака ($c \in I$) по возрастанию

$$r_{c_1}, r_{c_2}, ..., r_{c_m}. \tag{1}$$

Согласно данному ниже критерию последовательность (1) разбивается на τ_c ($\tau_c \geq 2$) непересекающихся интервалов $[r_{c_u}, r_{c_v}]^i$, $1 \leq u$, $u \leq v \leq m$, $i = \overline{1,\tau_c}$. Значения,

лежащие в интервале $[r_{c_u}, r_{c_v}]^i$, далее могут рассматриваться как градация номинального признака.

Пусть $d_t^i(u,v), \overline{d_t^i(u,v)}$ — количество представителей соответственно классов K_t и CK_t в интервале $[r_{c_u}, r_{c_v}]^i$. Для рекурсивной процедуры выбора значений r_{c_u}, r_{c_v} используется критерий

$$\left| \frac{d_t^i(u,v)}{\mid E_0 \cap K_t \mid} - \frac{\overline{d_t^i(u,v)}}{\mid E_0 \cap CK_t \mid} \right| \to \max. \tag{2}$$

Границы первого интервала $[r_{c_u}, r_{c_v}]^1$ на последовательности (1) вычисляются по максимуму критерия (2). Аналогичным образом определяются границы для $[r_{c_u}, r_{c_v}]^p$, $p > 1$, на значениях (1), не вошедших в $[r_{c_u}, r_{c_v}]^1, ..., [r_{c_u}, r_{c_v}]^{p-1}$. Критерием останова процедуры является покрытие всех значений (1) непересекающимися интервалами.

Обозначим через $\eta_{1i}(t) = \dfrac{d_t^i(u,v)}{\mid E_0 \cap K_t \mid}$, $\eta_{2i}(t) = \dfrac{\overline{d_t^i(u,v)}}{\mid E_0 \cap CK_t \mid}$ результаты оптимального разбиения по (2) для каждого интервала $[r_{c_u}, r_{c_v}]^i$, $i = \overline{1, \tau_c}$. Значение функции принадлежности c-го признака к K_t по интервалу $[r_{c_u}, r_{c_v}]^i$ определим как

$$f_{ci}(t) = \frac{\eta_{1i}(t)}{\eta_{1i}(t) + \eta_{2i}(t)}. \tag{3}$$

Если признак $c \in J$, то $\eta_{1i}(t), \eta_{2i}(t)$ в (3) рассматриваются как количество значений i-й градации у объектов E_0 соответственно из классов K_t и CK_t. Считается, что множество чисел, идентифицирующих τ_c градаций номинального признака, всегда можно взаимнооднозначно отобразить в множество $\{1, 2, ..., \tau_c\}$.

Будем использовать функции принадлежности (3) для отображения значений признаков в описании объектов E_0 на числовую ось. Предполагается, что по результатам отображения можно провести границу между объектами из K_d и CK_d. Обобщенная оценка объекта $S \in E_0$, $S = (b_1, b_2, ..., b_n)$, по классу K_d вычисляется по формуле

$$R(S) = \frac{1}{\mid T \mid} \sum_{S_j \in T} \left(\sum_{c \in I} \left\{ \begin{array}{ll} f_{ci}(d), b_c \in [r_{c_u}, r_{c_v}]^i \text{ и } x_{jc} \notin [r_{c_u}, r_{c_v}]^i & \\ \dfrac{f_{ci}(d) \mid b_c - x_{jc} \mid}{\mid r_{c_u} - r_{c_v} \mid}, & r_{c_u} \neq r_{c_v} \\ 0, & r_{c_u} = r_{c_v} \end{array} \right\}, \quad b_c, x_{jc} \in [r_{c_u}, r_{c_v}]^i \right. \; + \\ \left. + \sum_{c \in J} \left\{ \begin{array}{ll} f_{ci}(d), & b_c \neq x_{jc} \\ 0, & \text{в противном случае} \end{array} \right. \right), \tag{4}$$

где $S_j = (x_{j1}, x_{j2}, ..., x_{jn})$, $T = \left\{ \begin{array}{ll} E_0 \cap CK_d, & S \in K_d \\ E_0 \cap K_d, & S \in CK_d \end{array} \right.$ и значения τ_c градаций, $c \in J$, каждого номинального признака принадлежат множеству $\{1, 2, ..., \tau_c\}$.

Значения, вычисляемые по (4), являются средством упорядочения объектов по отношению к определяемому классу K_t, $t = \overline{1, l}$. Предметом изучения может быть изменение порядка следования объектов в зависимости от разных наборов признаков.

Разбиение на интервалы по (2) эксперты могут использовать при формировании лингвистических правил для баз знаний. Количество интервалов доминирования классов косвенно указывает на статус закономерностей. Чем меньше интервалов доминирования, тем сильнее проявление закономерности на конкретном признаке в классе. Этим

свойством можно пользоваться при ранжировании количественных показателей в прикладных задачах. Самые высокие ранги получают те показатели, число интервалов доминирования классов K_t и CK_t которых минимально. Дополнительной альтернативой для ранжирования по классу K_t при равном количестве интервалов служит показатель

$$g_c(t) = \frac{1}{m} \sum_{\{[r_{c_u}, r_{c_v}]^i\}} \begin{cases} f_{ci}(t)(v-u+1), & f_{ci}(t) > 0.5, \\ (1-f_{ci}(t))(v-u+1), & f_{ci}(t) < 0.5, \end{cases} \tag{5}$$

выражающий степень однородности (неперемешанности) значений c-го признака объектов из K_d и CK_d в границах интервалов доминирования, определяемых по (2).

2. Устойчивость объектов в разнотипном признаковом пространстве

Обозначим через Ω упорядоченное множество значений c-го признака ($c \in I$), равное (1). Специфика процесса вычисления устойчивости объектов классов с описанием в разнотипном признаковом пространстве такова: требуется определить разбиение на интервалы, в границах которых доминируют представители только одного класса. Этим целям служит критерий

$$\frac{d_t(u,v)}{\mid E_0 \cap K_t \mid} - \frac{\overline{d_t(u,v)}}{\mid E_0 \cap CK_t \mid} \to \max_{\Omega}, \tag{6}$$

где $d_t(u,v), \overline{d_t(u,v)}$ — количество значений c-го признака в $[r_{c_u}, r_{c_v}]_t$ соответственно из классов K_t и CK_t, $t = \overline{1,l}$.

Алгоритм выбора интервалов c-го признака с доминированием значений только одного из классов $K_1, K_2, ..., K_l$, $l > 2$, для вычисления устойчивости объектов реализуется следующим образом.

1. Выбор интервалов доминирования $\{[r_{c_u}, r_{c_v}]_t\}$, $t = \overline{1,l}$, каждого класса K_t относительно CK_t по (6). Вычисление $\eta_1(t) = \dfrac{d_t(u,v)}{\mid E_0 \cap K_t \mid}$, $\eta_2(t) = \dfrac{\overline{d_t(u,v)}}{\mid E_0 \cap CK_t \mid}$.

2. Вычисление значения функции принадлежности к классу K_t по $[r_{c_u}, r_{c_v}]_t$ с помощью формулы $z_c(t) = \dfrac{\eta_1(t)}{\eta_1(t) + \eta_2(t)}$.

3. Если $z_c(k) = \max\limits_{1 \le t \le l} z_c(t)$, то выбор в качестве интервала $[r_{c_u}, r_{c_v}]^i = [r_{c_u}, r_{c_v}]_k$, $i = 1, 2, ...$ и $\Omega = \Omega \setminus \{r_{c_d} \mid r_{c_d} \in [r_{c_u}, r_{c_v}]^i\}$.

4. Вычисление по (3) значения $f_{ci}(t)$, $t = \overline{1,l}$, функции принадлежности к классу K_t по интервалу $[r_{c_u}, r_{c_v}]^i$.

5. Если $\Omega \neq 0$, то переход на 1.

Разбиение на интервалы, в границах которых доминируют представители только одного класса и соответствующие им (интервалам) значения функций принадлежности (3), предлагается использовать для вычисления степени компактности (устойчивости) размещения объекта $S \in E_0$ относительно объектов непересекающихся классов $K_1, ..., K_l$.

Устойчивость объекта $S_i \in E_0$ ($S_i = (x_{i1}, x_{i2}, ..., x_{in})$) в классе K_t определяется по формуле

$$\gamma_t(S_i) = \frac{1}{n(\mid E_0 \cap K_t \mid -1)} \left(\sum_{c \in I,\, x_{ic} \in [r_{c_u}, r_{c_v}]^p} f_{cp}(t)(q_{tp} - 1) + \sum_{c \in J,\, x_{ic} = p} f_{cp}(t)(h_{tp} - 1) \right), \quad (7)$$

где q_{tp} — количество значений c-го признака из класса K_t в интервале $[r_{c_u}, r_{c_v}]^p$, h_{tp} — число значений градации $x_{ic} = p$ в описании объектов из $E_0 \cap K_t$. Множество значений, вычисляемых по (7), принадлежит $[0, 1]$ и может быть использовано для интерпретации устойчивости объектов в терминах нечеткой логики. Интерес для экспертов-исследователей представляет и среднее значение по устойчивости объектов классов $K_1, K_2, ..., K_l$.

Имеет смысл проверка утверждения о корреляции между устойчивостью объектов классов и числом объектов-эталонов локально-оптимального покрытия обучающей выборки в моделях ИНС с минимальной конфигурацией [3]. Заслуживает внимания изучение наличия связи между устойчивостью и способностью к обобщению нейронной сети. Действительно ли существует закономерность: чем меньше устойчивость объектов в некоторой области разнотипного признакового пространства, тем меньшей способностью к обобщению в этой области обладает нейронная сеть?

3. Вычислительный эксперимент

Для вычислительного эксперимента была взята выборка данных ИРИС Фишера [4], представленная 75 объектами. Выборка разделена на три непересекающихся класса по 25 объектов в каждом. Для описания допустимых объектов используются четыре количественных признака. Число интервалов доминирования значений признаков из классов K_t и CK_t, $t = \overline{1,3}$, по (2) представлено в табл. 1.

Порядок следования признаков по их значимости в классах приведен в табл. 2. В скобках указаны альтернативные значения (5) для признаков с равным количеством интервалов по (2). Устойчивость ряда объектов классов по (7) представлена в табл. 3.

Выбор локально-оптимального покрытия множества E_0 объектами-эталонами с помощью двух схем процедуры последовательного исключения [3] приведен в табл. 4. Каждое из двух полученных множеств объектов-эталонов покрытия позволяет корректно распознавать объекты обучающей выборки.

Из таблиц 3, 4 легко прослеживается коррелированность числа объектов-эталонов локально-оптимального покрытия с устойчивостью объектов классов. Число объектов-эталонов покрытия с относительно малой устойчивостью больше аналогичного числа

Таблица 1. Число интервалов по (2)

Номер класса	Номер признака			
	1	2	3	4
1	2	2	2	2
2	3	4	3	3
3	2	5	2	2

Таблица 2. Порядок следования признаков в классах

Класс	Порядок следования признаков
1	3 (1.0), 4 (1.0), 1 (0.896), 2 (0.799)
2	3 (0.979), 4 (0.968), 1 (0.754), 2
3	4 (0.967), 3 (0.960), 1 (0.787), 2

Т а б л и ц а 3. Устойчивость объектов классов

Номер объекта	Класс			Номер объекта	Класс		
	1	2	3		1	2	3
1 (1 класс)	0.8579	0.0043	0.0020	50 (2 класс)	0.0052	0.6726	0.0899
25 (1 класс)	0.8579	0.0043	0.0020	54 (3 класс)	0.0000	0.1601	0.6902
27 (2 класс)	0.0032	0.5162	0.1776	57 (3 класс)	0.2005	0.3467	0.3086
29 (2 класс)	0.0052	0.6726	0.0899	70 (3 класс)	0.0052	0.4498	0.3199
44 (2 класс)	0.0052	0.6726	0.0899	72 (3 класс)	0.0052	0.2151	0.5546
45 (2 класс)	0.0052	0.6726	0.0899	74 (3 класс)	0.0000	0.1601	0.6902
46 (2 класс)	0.0084	0.3365	0.2768	75 (3 класс)	0.0032	0.0587	0.6423

Т а б л и ц а 4. Объекты-эталоны покрытия

Схема исключения	Класс		
	1	2	3
$1,2,..,75$	25	$44,45,46,50$	$70,72,74,75$
$75,74,..,1$	1	$27,29,44$	$54,57,70$

для объектов-эталонов с относительно большой устойчивостью. К сожалению, нет формального обоснования для использования устойчивости объектов при решения задачи минимального покрытия в качестве альтернативы полному перебору.

С целью демонстрации вычисления устойчивости объектов, описываемых разнотипными признаками, воспользуемся результатами оптимизации критерия из [3] для отображения значений количественных признаков в номинальные. Идея реализации критерия такова. Упорядоченное множество значений признака $x_j,\ j \in I$, разбивается на ряд интервалов $(c_{2k-1}, c_{2k}]$, $c_{2k-1} < c_{2k},\ k = \overline{1,l}$, каждый из которых считается градацией номинального признака. Определение границ интервалов $(c_{2k-1}, c_{2k}]$ основано на проверке гипотезы (утверждения) о том, что каждый интервал содержит значения количественного признака объектов только одного класса.

Пусть u_i^p — число значений признака $x_j,\ j \in I$, из класса K_i в интервале $(c_{2p-1}, c_{2p}]$, $A = (a_0, ..., a_l)$, $a_0 = 0,\ a_l = m$, a_p — порядковый номер элемента упорядоченной по возрастанию последовательности $r_{j_1}, ..., r_{j_m}$ значений x_j у объектов из E_0, определяющий правую границу интервала $c_{2p} = r_{a_p}$. Критерий

$$\left(\frac{\sum\limits_{p=1}^{l} \sum\limits_{i=1}^{l} u_i^p (u_i^p - 1)}{\sum\limits_{i=1}^{l} \mid E_0 \cap K_i \mid (\mid E_0 \cap K_i \mid -1)} \right) \times$$

$$\times \left(\frac{\sum\limits_{p=1}^{l} \sum\limits_{i=1}^{l} u_i^p (m - \mid E_0 \cap K_i \mid - \sum\limits_{j=1}^{l} u_j^p + u_i^p)}{\sum\limits_{i=1}^{l} \mid E_0 \cap K_i \mid (m - \mid E_0 \cap K_i \mid)} \right) \to \max_{\{A\}} \tag{8}$$

позволяет вычислять оптимальные значения границ интервалов $\{(c_{2p-1}, c_{2p}]\}$ и использовать их для определения градаций количественного признака в номинальной шкале измерений. Если в границах каждого из l интервалов размещаются значения признака

Т а б л и ц а 5. Устойчивость объектов, описываемых разнотипными признаками

Номер объекта	Класс			Номер объекта	Класс		
	1	2	3		1	2	3
1 (1 класс)	0.8579	0.0043	0.0020	50 (2 класс)	0.0173	0.7371	0.2796
25 (1 класс)	0.8579	0.0043	0.0020	54 (3 класс)	0.0173	0.2796	0.7443
27 (2 класс)	0.0173	0.7371	0.2796	57 (3 класс)	0.2154	0.3799	0.3546
29 (2 класс)	0.0148	0.6103	0.1199	70 (3 класс)	0.0173	0.5143	0.5096
44 (2 класс)	0.0173	0.7371	0.2796	72 (3 класс)	0.0173	0.2796	0.7443
45 (2 класс)	0.0173	0.7371	0.2796	74 (3 класс)	0.0173	0.2796	0.7443
46 (2 класс)	0.0173	0.5024	0.5143	75 (3 класс)	0.0025	0.1283	0.6314

x_j, $j \in I$, объектов только одного из классов, то критерий (8) принимает значение, равное единице. Во всех остальных (не идеальных) случаях максимум критерия (8) принимает значение из интервала $(0, 1)$.

Устойчивость по (7) ряда объектов классов, описываемых двумя количественными и двумя номинальными признаками, показана в табл. 5. Для 1-го и 2-го признаков в описании объектов использовались номера интервалов (вычисленные по (8)), в границах которых лежали их исходные значения.

Слабое различие между значением устойчивости объекта-эталона покрытия из класса K_t и аналогичными значениями по $K_1, ..., K_{t-1}$, $K_{t+1}, ..., K_l$ указывает на ее плохую способность к обобщению в ИНС.

Заключение

Процесс вычисления устойчивости объектов классов является инвариантным относительно масштабов измерений количественных признаков. Свойство инвариантности существенно повышает возможности сравнения результатов на экспериментальных данных, полученных независимыми исследователями. Значения устойчивости, вычисленные по (7), позволяют исследовать структуры размещения объектов при различных сочетаниях разнотипных признаков, используемых для их описания. Такое исследование востребовано для проверки гипотезы о компактности, согласно которой производится разбиение объектов на классы.

Технологию вычисления устойчивости классифицированных объектов можно рекомендовать для моделирования процессов гомеостаза в различных предметных областях (например, медицине, геологии, биологии).

Список литературы

[1] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности: Справ. изд. М.: Финансы и статистика, 1989.

[2] Ignat'ev N.A., Adilova F.T., Matlatipov G.R., Chernysh P.P. Knowledge discovering from clinical data based on classification tasks solving // MediNFO. Amsterdam: IOS Press, 2001. P. 1354–1358.

[3] Игнатьев Н.А., Мадрахимов Ш.Ф. О некоторых способах повышения прозрачности нейронных сетей // Вычисл. технологии. 2003. Т. 8, № 6. С. 31–37.

[4] Wold S. Pattern recognition by means of disjoint principal components models // Patt. Recognit. 1976. Vol. 8, No. 3. P. 127–139.

Поступила в редакцию 11 января 2010 г.,
с доработки — 31 января 2011 г.