О конструировании признакового пространства для поиска логических закономерностей в задачах распознавания образов

Н. А. Игнатьев
Национальный университет Узбекистана, Ташкент
e-mail n_ignatev@rambler.ru

Рассматривается отображение представления объектов классов в разнотипном признаковом пространстве на числовые шкалы. Результаты отображения используются для поиска логических закономерностей в данных.

Ключевые слова: искусственный интеллект, обобщённые оценки, визуализация данных, логические закономерности, интеллектуальный анализ данных.

Введение

Поиск закономерностей в базах данных — одно из важнейших направлений интеллектуального анализа данных. Решение проблемы комбинаторной сложности такого поиска в логических методах обнаружения закономерностей в ранних работах сводилось к проблеме выбора вариантов за приемлемое время. Используемые для этих целей алгоритмы ограниченного перебора производили вычисление частоты комбинаций логических событий в подгруппах данных. Полезность той или иной комбинации определялась на основании анализа этих частот [1]. Рассматривались и альтернативные методы, позволяющие практически отказаться от перебора вариантов при поиске закономерностей.

Причиной отказа от перебора вариантов в [2] было утверждение о том, что в каждой точке (описании объекта) признакового пространства существует своя закономерность. В окрестности объекта конструировалось собственное пространство признаков и определялась индивидуальная мера его сходства с другими объектами. При конструировании использовался геометрический подход, основным видом операций которого являлась операция отображения описаний объектов на числовую шкалу. Для обнаружения логических закономерностей применялись средства линейной алгебры и интерактивной графики. При исследовании структуры множества логических закономерностей на основе геометрических представлений использовались методы визуализации данных.

Понижение размерности признакового пространства методами главных компонент и факторного анализа рассматривалось в [3]. Выбор тех или иных критериев для обоснования используемых методов визуализации многомерных данных основывался на эвристических соображениях. Результаты визуализации в основном применялись для разведочного анализа данных.

В настоящей работе новое признаковое пространство объектов предлагается строить с использованием методов вычисления обобщённых оценок [4]. Обобщённая оценка объекта представляет собой интегрированный количественный показатель по значениям определяемого множества его признаков. Результаты отображения обобщённых оценок

в R^1 используются для построения if-then правил с помощью логических закономерностей в форме полуплоскостей. При отображении в R^2 по визуальному представлению объектов можно проводить фильтрацию обучающей выборки. Новое (двумерное) признаковое пространство доступно для поиска всех известных форм логических закономерностей.

Существенное отличие предлагаемого метода визуализации от ранее известных [2, 3] заключается в следующем:

— на две числовые шкалы отображаются объекты с описанием в разнотипном признаковом пространстве;

— процесс отображения в новое (двумерное) признаковое пространство реализуется через вычисление обобщённых оценок.

1. Вычисление обобщённых оценок объектов

Рассматривается задача распознавания в стандартной постановке. Считается, что задано множество $E_0 = \{S_1, ..., S_m\}$ объектов, разделённое на два непересекающихся подмножества (класса) K_1, K_2. Описание объектов производится с помощью n разнотипных признаков, ξ из которых измеряются в интервальных шкалах, $(n-\xi)$ — в номинальной.

Отображение объектов E_0 на числовую шкалу производится функционалом $F(S, \Omega)$, где Ω — множество параметров, $S \in E_0$. Требуется определить значения параметров Ω, при которых

$$\min_{S \in K_1} F(S, \Omega) - \max_{S \in K_2} F(S, \Omega) \to \max .$$

Обозначим через I, J множество номеров соответственно количественных и номинальных (качественных) признаков $X = \{x_1, ..., x_n\}$ в описании допустимых объектов, $|I| + |J| = n$. Определим веса количественных признаков с учётом разделения объектов на классы K_1 и K_2.

Упорядоченное множество значений признака $x_j, j \in I$, разделим на два интервала $[c_1, c_2], (c_2, c_3]$, каждый из которых рассматривается как градация номинального признака. Критерий для определения границы c_2 основывается на проверке гипотезы (утверждения) о том, что каждый из двух интервалов содержит значения количественного признака объектов только одного класса.

Пусть u_i^1, u_i^2 — количество значений признака $x_j, j \in I$, класса K_i, $i = 1, 2$, соответственно в интервалах $[c_1, c_2], (c_2, c_3]$, $|K_i| > 1$, p — порядковый номер элемента упорядоченной по возрастанию последовательности $r_{j_1}, ..., r_{j_p}, ..., r_{j_m}$ значений x_j из E_0, определяющий границы интервалов как $c_1 = r_{j_1}, c_2 = r_{j_p}, c_3 = r_{j_m}$. Критерий

$$\left(\frac{\sum\limits_{i=1}^{2} u_i^1(u_i^1 - 1) + u_i^2(u_i^2 - 1)}{\sum\limits_{i=1}^{2} |K_i|(|K_i| - 1)} \right) \left(\frac{\sum\limits_{d=1}^{2} \sum\limits_{i=1}^{2} u_i^d \left(|K_{3-i}| - u_{3-i}^d \right)}{2|K_1||K_2|} \right) \to \max \qquad (1)$$

позволяет вычислять оптимальное значение границы между интервалами $[c_1, c_2], (c_2, c_3]$ и использовать её для определения градаций количественного признака в номинальной шкале измерений. Выражение в левых скобках (1) представляет внутриклассовое сходство, в правых — межклассовое различие.

Пусть w_i — оптимальное значение критерия (1) по i-му ($i \in I$) признаку ($0 < w_i \leq 1$), c_1^i, c_2^i, c_3^i — соответствующие этому значению концы интервалов разбиения $[c_1^i, c_2^i], (c_2^i, c_3^i]$. Для вычисления обобщённой оценки произвольного допустимого объекта $S = (x_1, ..., x_n)$, все признаки которого количественные, используется функционал

$$R(S) = \sum_{i=1}^{n} w_i t_i (x_i - c_2^i)/(c_3^i - c_1^i),$$

где значения элементов вектора $T = (t_1, ..., t_n)$, $t_i \in \{-1, 1\}$, определяются из условия

$$\min_{S_p \in K_1} R(S_p) - \max_{S_p \in K_2} R(S_p) \to \max. \tag{2}$$

Поиск решения многоэкстремальной задачи по (2) производится алгоритмом стохастической оптимизации. Пошаговая реализация этого алгоритма следующая.

1. Выбор числа итераций k, $\left[\frac{n}{2}\right] \leq k \leq n$, iter $= 0$, $T_{\max} = (\overbrace{1, \ldots, 1}^{n})$, $Z_{\max} = -n$.

2. iter $=$ iter $+ 1$. $\Omega = \{1, \ldots, n\}$. Выбор начального значения вектора $T = (t_1, \ldots, t_n)$ для новой (iter-й по счёту) итерации. Вычисление значений $R(S_j)$ ($S_j = (x_{j1}, \ldots, x_{jn})$) $\forall\ S_j \in E_0$ и

$$Z = \min_{S_j \in K_1} R(S_j) - \max_{S_j \in K_2} R(S_j).$$

3. $\forall\ i \in \Omega$ вычисление значения (при условии замены t_i на $-t_i$)

$$Z_i = \min_{S_j \in K_1} R^*(S_j) - \max_{S_j \in K_2} R^*(S_j),$$

где

$$R^*(S_j) = R(S_j) - 2t_i w_i (x_{ji} - c_2^i)/(c_3^i - c_1^i), j = \overline{1, m}.$$

4. $Z_p = \max\limits_{i \in \Omega} Z_i$. Если $Z_p > Z$, то $Z = Z_p$, $\Omega = \Omega \backslash p$,

$$R(S_j) = R(S_j) - 2t_p w_p (x_{jp} - c_2^p)/(c_3^p - c_1^p), \quad j = \overline{1, m}, \quad t_p = -t_p,$$

и переход на шаг 3.

5. Если $Z > Z_{\max}$ то $Z_{\max} = Z$, $T_{\max} = T$.

6. Если iter $< k$, то 2.

7. Вывод $Z_{\max}$, $T_{\max}$.

Шаги алгоритма 2 — 4 представляют вычисление локальных максимумов при разных начальных значениях элементов вектора T. Максимальное значение $Z_{\max}$ среди локальных максимумов и соответствующие ему значения элементов вектора $T_{\max}$ (шаг 5) выбираются в качестве решения задачи по условию (2).

Для вычисления обобщённых оценок объектов с описанием в разнотипном признаковом пространстве дополнительно требуется определять значения весов номинальных признаков и вкладов их градаций.

Введем обозначения: p — число градаций признака $r \in J$, g_{dr}^t — число значений t-й ($1 \leq t \leq p$) градации r-го признака в описании объектов класса K_d, l_{dr} — число градаций r-го признака в $K_d, d = 1, 2$. Различие по r-му признаку между классами K_1 и K_2 определяется как величина

$$\lambda_r = 1 - \frac{\sum\limits_{t=1}^{p} g_{1r}^t g_{2r}^t}{|K_1||K_2|}. \tag{3}$$

Степень однородности (мера внутриклассового сходства) β_r значений градаций r-го признака по классам K_1, K_2 вычисляется по формулам

$$D_{dr} = \begin{cases} (|K_d| - l_{dr} + 1)(|K_d| - l_{dr}), p > 2, \\ |K_d|(|K_d| - 1), p \leq 2, \end{cases}$$

$$\beta_r = \begin{cases} \dfrac{\sum\limits_{t=1}^{p} g_{1r}^t(g_{1r}^t - 1) + g_{2r}^t(g_{2r}^t - 1)}{D_{1r} + D_{2r}}, \; D_{1r} + D_{2r} > 0, \\ 0, D_{1r} + D_{2r} = 0. \end{cases} \tag{4}$$

С помощью (3),(4) вес номинального признака $r \in J$ определяется как

$$v_r = \lambda_r \beta_r.$$

Очевидно, что множество чисел, идентифицирующих p градаций номинального признака, всегда можно взаимно однозначно отобразить в множество $\{1, ..., p\}$. С учётом такого отображения для объекта $S = (x_1, ..., x_n)$ вклад признака $x_i = j$, $i \in J$, $j \in \{1, ..., p\}$, в обобщённую оценку определяется величиной

$$\mu_i(j) = v_i \left(\frac{\alpha_{ij}^1}{|K_1|} - \frac{\alpha_{ij}^2}{|K_2|} \right), \tag{5}$$

где $\alpha_{ij}^1, \alpha_{ij}^2$ — количество значений j-й градации i-го признака соответственно в классах K_1 и K_2, v_i — вес i-го признака. При наличии показателей, измеряемых в номинальной шкале, обобщённая оценка для каждого объекта $S_a \in E_0$, $S_a = (x_{a1}, ..., x_{an})$ будет вычисляться как

$$R(S_a) = \sum_{i \in I} w_i t_i (x_{ai} - c_2^i)/(c_3^i - c_1^i) + \sum_{i \in J} \mu_i(x_{ai}). \tag{6}$$

2. Представление объектов в новом (двумерном) признаковом пространстве

Целью конструирования нового признакового пространства является визуализация объектов, описываемых разнотипными признаками. В работе [2] максимальное сохранение структурных особенностей размещения объектов при отображении в двумерное признаковое пространство основывалось на применении методов линейной алгебры. В настоящей работе для аналогичных целей используются функционалы, значения параметров которых вычисляются по определяемым локальным областям признакового пространства.

Выбор двух числовых шкал для отображения на них представления объектов в разнотипном признаковом пространстве функционалами R_1, R_2 производится следующим образом. Для вычисления параметров функционалов R_1, R_2 исходное признаковое пространство делится на три локальные области L_1, L_2, L_3, что предусматривает как пропорциональное представительство в них объектов из K_1, K_2, так и учёт структуры размещения объектов в выборке.

По обобщённым оценкам (6) строится упорядоченная по убыванию последовательность $S_{i_1}, ..., S_{i_m}$ объектов E_0. В область L_1 включаются объекты последовательности

с S_{i_1} по $S_{i_{[K_1/2]}}$, в L_3 с $S_{i_{m-[K_2/2]}}$ по S_{i_m}. Объекты, не вошедшие в области L_1, L_3, попадают в область L_2.

Для первой числовой шкалы множество параметров $\{w_i\}, \{c_2^i\}, i \in I$, и вкладов градаций $\{\mu_i(j)\}, i \in J$, по (1)–(5) функционала R_1 вычисляется в $L_1 \cup L_3$. Аналогичные вычисления для второй числовой шкалы (по функционалу R_2) производятся в L_2. В целях сохранения масштаба измерений значения $\{c_1^i\}\{c_3^i\}, i \in I$, остаются неизменными при отображении объектов на числовые шкалы всеми функционалами типа (6).

Проверка равносильности числовых шкал для отображения на них объектов E_0 функционалами R_1 и R_2 производится с помощью критерия (1). Под равносильностью понимается сохранение масштаба измерений и стабильности структуры взаимного размещения объектов в новом признаковом пространстве. Выражением (показателем) равносильности служит максимальная близость значений критерия (1) по обобщённым оценкам объектов E_0, полученным по R_1 и R_2.

Новое (двумерное) признаковое пространство является результатом предобработки данных, используя которые можно вычислять обобщённые оценки как описанным выше методом, так и методом с применением интервалов доминирования количественных признаков [4]. Привлекательность последнего метода заключается в гарантированной однозначности значений полученных оценок на E_0. Эти оценки могут быть использованы в качестве значения целевого признака при построении функции регрессии в новом признаковом пространстве.

Рассмотрим выбор интервалов количественного признака в двумерном пространстве, в границах которых доминируют значения из класса K_1 или K_2. Для этого упорядочим значения c-го признака, $c = 1, 2$, объектов E_0 по возрастанию

$$r_{c_1}, r_{c_2}, ..., r_{c_m}. \tag{7}$$

Согласно определяемому ниже критерия последовательность (7) разбивается на τ_c ($\tau_c \geq 2$) непересекающихся интервалов $[r_{c_u}, r_{c_v}]^i$, $1 \leq u, u \leq v \leq m$, $i = \overline{1, \tau_c}$.

Пусть $d_1^i(u, v), d_2^i(u, v)$ — количество представителей соответственно классов K_1, K_2 в интервале $[r_{c_u}, r_{c_v}]^i$. Для рекурсивной процедуры выбора значений r_{c_u}, r_{c_v} используется критерий

$$\left| \frac{d_t^i(u, v)}{|K_1|} - \frac{d_t^i(u, v)}{|K_2|} \right| \to \max. \tag{8}$$

Границы первого интервала $[r_{c_u}, r_{c_v}]^1$ на последовательности (7) вычисляются по максимуму критерия (8). Аналогичным образом определяются границы для $[r_{c_u}, r_{c_v}]^p$, $p > 1$, на значениях (7), не вошедших в $[r_{c_u}, r_{c_v}]^1, ..., [r_{c_u}, r_{c_v}]^{p-1}$. Критерием останова процедуры служит покрытие всех значений (7) непересекающимися интервалами.

Обозначим через $\eta_{1i} = \dfrac{d_1^i(u, v)}{|K_1|}$, $\eta_{2i} = \dfrac{d_2^i(u, v)}{|K_2|}$ результаты оптимального разбиения по (8) для каждого интервала $[r_{c_u}, r_{c_v}]^i$, $i = \overline{1, \tau_c}$. Значение функции принадлежности c-го признака к K_1 по интервалу $[r_{c_u}, r_{c_v}]^i$ определим как

$$f_{ci} = \frac{\eta_{1i}}{\eta_{1i} + \eta_{2i}}.$$

Обобщённая оценка объекта $S \in E_0$, $S = (b_1, b_2)$ вычисляется по формуле

$$Q(S) = \frac{1}{|Z|} \sum_{S_j \in Z} \sum_{c=1}^{2} \left\{ \begin{array}{l} f_{ci}, b_c \in [r_{c_u}, r_{c_v}]^i \text{ и } x_{jc} \notin [r_{c_u}, r_{c_v}]^i, \\ \dfrac{f_{ci}|b_c - x_{jc}|}{|r_{c_u} - r_{c_v}|}, \ r_{c_u} \neq r_{c_v}, \\ 0, \ r_{c_u} = r_{c_v} \end{array} \right\}, \quad b_c, x_{jc} \in [r_{c_u}, r_{c_v}]^i, \tag{9}$$

где

$$S_j = (x_{j1}, x_{j2}), \quad Z = \left\{ \begin{array}{ll} E_0 \cap K_2, & S \in K_1 \\ E_0 \cap K_1, & S \in K_2. \end{array} \right.$$

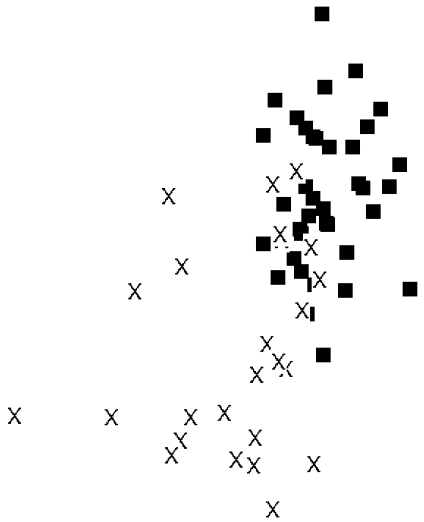
3. Вычислительный эксперимент

Для вычислительного эксперимента использовались данные из [5]. Множество E_0 представлялось как обучающая выборка из 66 объектов, содержащая значения показателей больных с различными степенями заболевания ишемической болезнью сердца (класс K_1) и практически здоровых людей (класс K_2). Объекты описывались 24 признаками, 14 из которых измерялись в количественных шкалах, 10 в номинальной.

Отображение выборки на числовую шкалу производилось функционалом $R(S)$, параметры которого вычислялись на E_0. Логическая закономерность в форме полуплоскости для первого класса определялась предикатом $\varphi_1(S, w) = [R(S) > r_2]$, $r_2 = \max\limits_{S \in K_2} R(S) = 0.4002$, для второго — $\varphi_2(S, w) = [R(S) < r_1]$, $r_1 = \min\limits_{S \in K_1} R(S) = 0.1016$.

Отрицательная величина $r_1 - r_2 = -0.3006$ по критерию (2) указывает на то, что точного разделения объектов классов на числовой шкале не произошло. Число ошибок для первого класса по $\varphi_1(S, w)$ было 18 из общего числа объектов 39, для второго класса по $\varphi_2(S, w)$ — соответственно 7 из 27.

Из значений обобщённых оценок 66 объектов по $R(S)$ была построена упорядоченная в порядке убывания последовательность. В локальную область L_1 вошли объекты этой последовательности с 1 по 19, в L_3 — соответственно с 53 по 66. Все оставшиеся объекты E_0 вошли в область L_2.



Отображение объектов на плоскость; × — объекты класса K_1, ■ — объекты класс K_2

Параметры $R_1(S)$ вычислялись по $L_1 \cup L_3$, $R_2(S)$ по L_2. Отображения объектов E_0 на две числовые шкалы по оценкам $R_1(S)$ и $R_2(S)$ имели значения критерия (1) соответственно 0.461 и 0.458. Результаты отображения показаны на рисунке.

Обобщённые оценки по (9) в двумерном пространстве имели следующие характеристики: $\min\limits_{S\in K_1} Q(S) - \max\limits_{S\in K_2} Q(S) = 0.0191$ при $\max\limits_{S\in K_1} Q(S) = 1.2837$ и $\min\limits_{S\in K_2} Q(S) = 0.2414$. Вычисление оценок по (6) в новом признаковом пространстве не дало хороших результатов. Так, максимальная разность по критерию (2) составила $\min\limits_{S\in K_1} R(S) - \max\limits_{S\in K_2} R(S) = -0.2026$. В рамках исследуемой задачи оценки по (9) могут служить значениями табличной функции для построения регрессии в новом признаковом пространстве.

Результаты конструирования признакового пространства описанными в данном исследовании методами востребованы в задачах интеллектуального анализа данных для построения моделей в слабо формализованных предметных областях. В новом признаковом пространстве могут быть использованы методы выделения различных форм логических закономерностей, проводится кластерный и регрессионный анализ данных.

Список литературы

[1] Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск, 1981.

[2] Дюк В.А. Формирование знаний в системах искусственного интеллекта: Геометрический подход // Вестник академии техн. творчества. 1996. №2. С. 46–67.

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

[4] Игнатьев Н.А. Вычисление обобщённых показателей и интеллектуальный анализ данных // Автоматика и телемеханика. 2011. №5. С. 183–190.

[5] Юлдашов Р.У. Интеллектуальный анализ данных в нейроэкспертных системах и задачи прогнозирования: Дис. ... канд. тех. наук. Ташкент: НУУз, 2011. 107 с.

Поступила в редакцию 30 марта 2012 г.,
с доработки — 22 июня 2012 г.