Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем

Н. А. Игнатьев

Национальный университет Узбекистана, Ташкент

*Контактный e-mail: n_ignatev@rambler.ru

Рассматривается метод разбиения обучающей выборки на непересекающиеся группы объектов на базе свойства связанности их по определяемому подмножеству граничных объектов классов. Разбиение на группы используется для поиска покрытия выборки объектами-эталонами. Описывается формирование нового признакового пространства для представления объектов путем нелинейного отображения непересекающихся наборов признаков на числовую ось.

Ключевые слова: распознавание образов, логические закономерности, кластерный анализ данных, оболочка классов, объекты-эталоны.

Введение

Основной целью кластерного анализа данных, представленных в обучающей выборке, является обоснование выбора и реализации алгоритмов распознавания. Для выбора моделей алгоритмов распознавания необходимо наличие информации [1] о различных структурах связей между объектами и признаками. В качестве одного из средств получения такой информации являются методы кластерного анализа данных. Структура связей между объектами классов зависит от используемой меры близости и преобразований признакового пространства. Наряду с различными способами нормирования данных к числу преобразований относятся формирование нового пространства на основе исходного и удаление неинформативных признаков.

Проблема формирования подмножеств информативных объектов и признаков, которые отражают закономерности обучающей выборки лучше, чем наборы исходных объектов и признаков, затрагивалась в [2]. Предлагалось для описания классов набором эталонных объектов (“столпов”) использовать функции конкурентного сходства (Frisфункции). Каждый столп защищал свою часть выборки — кластер, который определялся по значению функции конкурентного сходства относительно ближайшего объекта из противоположного класса.

Технология выбора столпов основана на оценке вклада в компактность классов каждого объекта. Значение вклада использовалось для селекции обучающих выборок с целью повышения обобщающей способности решающих правил. Для выбора информативных наборов признаков в [2] предложен алгоритм Frisgrad. Оценка качества обучающих выборок через компактность по системе столпов и наборов информативных

признаков не нашла отражения в демонстрации монотонности неубывания ее (оценки) значений при снижении размерности признакового пространства и невозрастании числа объектов-эталонов (столпов) классов.

На исследование структуры выборки с целью обнаружения логических закономерностей ориентирован геометрический подход на основе локальных метрик [3]. Отображения исходного признакового пространства в пространство размерности не выше k ($k \leq 3$) относительно определяемых экспертами центров позволяет получать визуальные представления отношений между объектами выборки. По результатам визуализации эксперты могут проводить селекцию выборки путем удаления шумовых объектов, делать выводы относительно устойчивости обнаруженных логических закономерностей. Анализ геометрической структуры данных методом локальной геометрии не имеет готовых шаблонов и реализуется известными методами и алгоритмами, использующими геометрическое описание данных [4].

Для кластерного анализа структуры обучающей выборки в данной работе используется подмножество граничных по заданной метрике объектов (оболочка) классов. Разбиение на непересекающиеся между собой группы объектов реализуется с помощью оболочки классов на основе свойства связанности. Согласно этому свойству для любых двух представителей группы существует цепочка (путь) из объектов, их соединяющих. Пара представителей определяет начало и конец цепочки, не выходящей за границы группы.

Свойство связанности гарантируют единственность решения на обучающей выборке, при котором число групп и их состав остаются неизменными. Предобработка данных с использованием разбиения на группы позволяет при выборе объектов-эталонов покрытия классов не прибегать к полному перебору всевозможных вариантов.

Другой целью кластерного анализа данных является снижение размерности признакового пространства путем формирования групп из непересекающихся наборов признаков и нелинейного отображения их значений в описании объектов на числовую ось. Экспериментальным путем доказывается, что набор объектов-эталонов классов в новом признаковом пространстве отражает логические закономерности лучше, чем в исходном пространстве.

1. О покрытии обучающей выборки объектами-эталонами

Рассматривается задача распознавания в стандартной постановке. Считается, что задано множество $E_0 = \{S_1, \ldots, S_m\}$ объектов, разделенное на $l (l \geq 2)$ непересекающихся подмножеств (классов) $K_1, \ldots, K_l$, $E_0 = \bigcup_{i=1}^{l} K_i$. Описание объектов производится с помощью набора из n разнотипных признаков $X(n) = (x_1, \ldots, x_n)$, ξ из которых измеряются в интервальных шкалах, $(n - \xi)$ — в номинальной. На множестве объектов E_0 задана метрика $\rho(x, y)$.

Обозначим через $L(E_0, \rho)$ подмножество граничных объектов классов, определяемое на E_0 по метрике $\rho(x, y)$. Объекты $S_i, S_j \in K_t$, $t = 1, \ldots, l$, считаются связанными между собой ($S_i \leftrightarrow S_j$), если

$$\{S \in L(E_0, \rho) \mid \rho(S, S_i) < r_i \text{ and } \rho(S, S_j) < r_j\} \neq \emptyset,$$

где $r_i(r_j)$ — расстояние до ближайшего от $S_i(S_j)$ объекта из CK_t ($CK_t = E_0 \backslash K_t$) по метрике $\rho(x, y)$. Множество $G_{t\nu} = \{S_{\nu_1}, \ldots, S_{\nu_c}\}$, $c \geq 2$, $G_{t\nu} \subset K_t$, $\nu \leq \mid K_t \mid$, представляет

область (группу) со связанными объектами в классе K_t, если для любых $S_{\nu_i}, S_{\nu_j} \in G_{t\nu}$ существует путь $S_{\nu_i} \leftrightarrow S_{\nu_k} \leftrightarrow \cdots \leftrightarrow S_{\nu_j}$. Требуется определить:

— минимальное число групп из связанных объектов по каждому классу K_t, $t = 1, \ldots, l$;

— минимальное покрытие множества E_0 объектами-эталонами для алгоритмов распознавания по прецедентам.

Минимальное число групп связанных объектов классов определяется на основе предобработки данных. Предобработка данных заключается:

— в выделении оболочки — подмножества граничных объектов классов $L(E_0, \rho)$ по заданной метрике ρ [5];

— описании объектов каждого класса по своей системе бинарных признаков.

Для выделения оболочки классов для каждого $S_i \in K_t$, $t = 1, \ldots, l$, построим упорядоченную по $\rho(x, y)$ последовательность

$$S_{i_0}, S_{i_1}, \ldots, S_{i_{m-1}}, S_i = S_{i_0}. \tag{1}$$

Пусть $S_{i_\beta} \in CK_t$ — ближайший к S_i объект из (1), не входящий в класс K_t. Обозначим через $O(S_i)$ окрестность радиуса $r_i = \rho(S_i, S_{i_\beta})$ с центром в S_i, включающую все объекты, для которых $\rho(S_i, S_{i_\tau}) < r_i, \tau = 1, \ldots, \beta - 1$. В $O(S_i)$ всегда существует непустое подмножество объектов

$$\Delta_i = \{S_{i_\alpha} \in O(S_i) \mid \rho(S_{i_\beta}, S_{i_\alpha}) = \min_{S_{i_\tau} \in O(S_i)} \rho(S_{i_\beta}, S_{i_\tau})\}. \tag{2}$$

По (2) принадлежность объектов к оболочке классов определяется как $L(E_0, \rho) = \bigcup_{i=1}^{m} \Delta_i$.

Множество объектов оболочки из $K_t \cap L(E_0, \rho)$ обозначим как $L_t(E_0, \rho) = \{S^1, \ldots, S^\pi\}$, $\pi \geq 1$. Значение $\pi = 1$ однозначно определяет вхождение всех объектов класса в одну область. При $\pi \geq 2$ преобразуем описание каждого объекта $S_i \in K_t$ в $S_i = (y_{i1}, \ldots, y_{i\pi})$, где

$$y_{ij} = \begin{cases} 1, & \rho(S_i, S^j) < r_i, \\ 0, & \rho(S_i, S^j) \geq r_i. \end{cases} \tag{3}$$

Пусть по (3) получено описание объектов класса K_t в новом (бинарном) признаковом пространстве, $\Omega = K_t$; θ — число не пересекающихся между собой групп объектов; $S_i \vee S_j$, $S_i \wedge S_j$ — соответственно операции дизъюнкции и конъюнкции по бинарным признакам объектов S_i, $S_j \in K_t$. Приведем пошаговое выполнение алгоритма разбиения объектов K_t на непересекающиеся группы $G_1, \ldots, G_\theta$.

Шаг 1. $\theta = 0$.

Шаг 2. Выделить объект $S \in \Omega$, $\theta = \theta + 1$, $Z = S$, $G_\theta = \emptyset$.

Шаг 3. **Выполнять** Выбор $S \in \Omega$ and $S \wedge Z = true$, $\Omega = \Omega \backslash S$, $G_\theta = G_\theta \cup S$, $Z = Z \vee S$, **пока** $\{S \in \Omega \mid S \wedge Z = true\} \neq \emptyset$.

Шаг 4. Если $\Omega \neq \emptyset$, то идти 2.

Шаг 5. Конец.

Поиск экстремума задачи о минимальном покрытии объектами-эталонами обучающей выборки связан с перебором множества различных вариантов. Все методы поиска, отличные от полного перебора, либо гарантируют локально-оптимальное решение задачи, либо основаны на использовании закономерностей, исключающих просмотр неперспективных вариантов. Разбиение на группы связанных между собой объектов классов

проводится с целью упорядочения процесса отбора объектов-эталонов минимального покрытия и цензурирования обучающей выборки.

Цензурирование обучающей выборки необходимо для определения обобщающей способности алгоритмов распознавания. Улучшение качества решающих правил возможно через селекцию объектов оболочки классов и обновление ее состава. Детальное исследование обобщающей способности распознающих алгоритмов в данной работе не рассматривается.

Обозначим через $R_S = \rho(S, \overline{S})$ расстояние от объекта $S \in K_t$ до ближайшего объекта $\overline{S}$ из противоположного к K_t класса ($\overline{S} \in CK_t$ и $S \neq \overline{S}$), через δ — минимальное число групп из связанных объектов в E_0. Для поиска минимального покрытия объектами-эталонами обучающей выборки упорядочим объекты каждой группы $G_u \cap K_t$, $u = 1, \ldots, \delta$, $t = 1, \ldots, l$, по множеству значений $\{R_S\}_{S \in G_u}$. В качестве меры близости между $S \in G_u$, $u = 1, \ldots, \delta$, и произвольным допустимым объектом S' используется взвешенное расстояние $d(S, S') = \rho(S, S')/R_S$. Решение о принадлежности S' к одному из классов $K_1, \ldots, K_l$ принимается по правилу: $S' \in K_t$, если

$$d(S_\mu, S') = \min_{S_u \in E_0} d(S_u, S') \text{ and } S_\mu \in K_t. \tag{4}$$

Согласно принципу *последовательного исключения*, используемому в процессе поиска покрытия, выборка E_0 делится на два подмножества: множество эталонов E_{ed} и контрольное множество E_k, $E_0 = E_{ed} \cup E_k$. В начале процесса $E_{ed} = E_0$, $E_k = \emptyset$. Упорядочение по значениям отступа $\{R_S\}_{S \in G_u}$, $u = 1, \ldots, \delta$, используется для определения кандидата на удаление из числа объектов-эталонов по группе G_u. Идея отбора заключается в поиске минимального числа эталонов, при котором алгоритм распознавания по (4) остается корректным (без ошибок распознающим объекты) на E_0.

Будем считать, что нумерация групп из связанных между собой объектов отражает порядок $\mid G_1 \mid \geq \cdots \geq \mid G_\delta \mid$ и по группе G_p, $p = 1, \ldots, \delta$, не производился отбор эталонных объектов. Кандидаты на удаление из E_{ed} последовательно выбираются начиная с $S \in G_p$ с минимальным значением R_S. Если включение $S \in E_k$ нарушает корректность решающего правила (4), то S возвращается во множество E_{ed}.

2. О нелинейном отображении наборов признаков на числовую ось

Предлагается метод формирования нового признакового пространства с использованием иерархической агломеративной группировки. С помощью этого метода производится нелинейное отображение множества значений из непересекающихся наборов признаков на числовую ось. Результаты отображения используются в качестве новых (латентных) признаков в описании объектов.

Правило для объединения признаков на каждом шаге иерархической группировки рассчитано на обучающую выборку с двумя непересекающимися классами, объекты которых описываются с помощью набора $X(n)$ из n количественных признаков. Для удобства изложения обозначим классы как A_1 и A_2, множество исходных номеров количественных признаков — как I, признаки, полученные на p-м шаге иерархической агломеративной группировки, — как $x_j^p, j \in i, p \geq 0$. При $p = 0$ $I = \{1, \ldots, n\}$. Если число классов $l \geq 3$, то к разбиению на два класса можно перейти, рассматривая объекты класса A_1 как $A_1 = K_t$, $t = 1, \ldots, l$, и A_2 — как $A_2 = CK_t$.

Упорядоченное множество значений признака $x_j^p, j \in I, p \geq 0$, объектов из E_0 разделим на два интервала $[c_1^{jp}, c_2^{jp}], (c_2^{jp}, c_3^{jp}]$, каждый из которых рассматривается как градация номинального признака. Критерий для определения границы c_2^{jp} основывается на проверке гипотезы (утверждения) о том, что каждый из двух интервалов содержит значения количественного признака объектов только из класса A_1 или A_2.

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

$$\left(\frac{\sum\limits_{i=1}^{2} u_i^1(u_i^1 - 1) + u_i^2(u_i^2 - 1)}{\sum\limits_{i=1}^{2} |A_i| (|A_i| - 1)} \right) \left(\frac{\sum\limits_{d=1}^{2} \sum\limits_{i=1}^{2} u_i^d(|A_{3-i}| - u_{3-i}^d)}{2 |A_1| |A_2|} \right) \to \max_{c_1^{jp} < c_2^{jp} < c_3^{jp}} \tag{5}$$

позволяет вычислять оптимальное значение границы между интервалами $[c_1^{jp}, c_2^{jp}]$ и $(c_2^{jp}, c_3^{jp}]$. Выражение в левых скобках (5) представляет внутриклассовое сходство, в правых — межклассовое различие.

Экстремум критерия (5) используется в качестве веса w_j^p $(0 \leq w_j^p \leq 1)$ признака x_j^p. При $w_j^p = 1$ значения признака x_j^p у объектов из классов A_1 и A_2 не пересекаются между собой.

Значение комбинации из двух признаков b_{rij}^p объекта $S_r = \{a_{ru}^p\}_{u \in I}$, $S_r \in E_0$, по паре (x_i^p, x_j^p), $0 \leq p < n$, $i, j \in I, i \neq j$, вычисляется как

$$b_{rij}^p = \eta_{ij} \left(\frac{t_i w_i^p (a_{ri}^p - c_2^{ip})}{(c_3^{ip} - c_1^{ip})} + \frac{t_j w_j^p (a_{rj}^p - c_2^{jp})}{(c_3^{jp} - c_1^{jp})} \right) +$$

$$+ \frac{(1 - \eta_{ij}) t_{ij} w_{ij}^p (a_{ri}^p a_{rj}^p - c_2^{ijp})}{(c_3^{ijp} - c_1^{ijp})}, \quad i, j \in i, \quad t_{ij}, t_i, t_j \in \{-1, 1\}, \quad \eta_{ij} \in [0, 1],$$

где w_i^p,w_j^p, w_{ij}^p — веса признаков, определяемые по (5) соответственно по множеству значений признаков x_i^p, x_j^p и их произведению $x_i^p x_j^p$; значения $t_{ij}, t_i, t_j \in \{-1, 1\}$, $\eta_{ij} \in [0, 1]$ выбираются по экстремуму функционала

$$\varphi(p, i, j) = \frac{\min\limits_{S_r \in K_1} b_{rij}^p - \max\limits_{S_r \in K_2} b_{rij}^p}{\max\limits_{S_r \in E_0} b_{rij}^p - \min\limits_{S_r \in E_0} b_{rij}^p} = \max_{t_{ij}, t_i, t_j \in \{-1,1\}, \eta_{ij} \in [0,1]}. \tag{6}$$

Экстремум функционала (6) интерпретируется как отступ между объектами классов A_1 и A_2 по множеству значений по паре признаков (x_i^p, x_j^p), $0 \leq p < n$, $i, j \in I$, $i \neq j$.

Обозначим через $\{z_{ij}^p\}_{i,j \in I}$, $p \geq 0$, квадратную матрицу размера $(n - p) \times (n - p)$, значение элемента z_{ij}^p которой при $p = 0$ определяется как

$$z_{ij}^p = \begin{cases} w_i^p, & i = j, \\ \text{значению (5) по } \{b_{rij}^p\}_{r=1}^m, & i \neq j, \end{cases} \tag{7}$$

через $\Gamma_\eta, \eta > 0$, — подмножество номеров признаков из $X(n)$. Приведем пошаговую реализацию алгоритма иерархической агломеративной группировки.

Шаг 1. $p = 0$, $\lambda c = 0$, $\eta = 1$. **Выполнять** $\Gamma_\eta = \{\eta\}$, $Margin_\eta = -2$, $\eta = \eta + 1$, **пока** $\eta \le n$.

Шаг 2. Вычислить значения элементов матрицы $\{z_{ij}^p\}_{i,j\in I}$ по (7).

Шаг 3. Выделить $\Phi = \{z_{uv}^p \mid z_{uv}^p \ge max(w_u^p, w_v^p) \text{ and } u \ne v, u, v \in I\}$. Если $\Phi = \emptyset$, то идти 9.

Шаг 4. Вычислить $\lambda n = \max\limits_{z_{u,v}^p \in \Phi} z_{uv}^p$. Выделить $\Delta = \{(s,t), s,t \in I \mid z_{st}^p = \lambda n \text{ and } s < t\}$. Определить пару $\{i,j\}$, $i < j$ как

$$\{i,j\} = \begin{cases} \Delta, & \mid \Delta \mid = 1, \\ \{s,t\}, & (s,t) \in \Delta \text{ and } \varphi(p,s,t) > \max\limits_{(u,v)\in\Delta\setminus(s,t)} \varphi(p,u,v). \end{cases}$$

Шаг 5. Если $\lambda n > \lambda c$ или $\lambda n = \lambda c$ и $Margin_i < \varphi(p,i,j)$, то $\gamma_i = \Gamma_i \cup \Gamma_j$, $\Gamma_j = \emptyset$, $Margin_i = \varphi(p,i,j)$, идти 7.

Шаг 6. Вывод номеров признаков из Γ_i, $\Gamma_i = \emptyset$, $I = I \setminus \{i\}$, идти 3.

Шаг 7. $p = p + 1$, $I = I \setminus max(i,j)$, $k = min(i,j)$, $\lambda c = \lambda n$. Заменить значения признаков в описании объекта $S_r = \{a_{ru}^{p-1}\}_{u\in I}$, $r = 1, \ldots, m$, на

$$a_{ru}^p = \begin{cases} a_{ru}^{p-1}, & u \in I \setminus k, \\ b_{rij}^p, & u = k. \end{cases}$$

Шаг 8. Для каждой пары (u,v), $u,v \in I$ определить значение

$$z_{uv}^p = \begin{cases} z_{uv}^{p-1}, & u \in I \setminus \{k\}, \quad v \in I, \\ \text{значению (5) на } \{a_{rv}^p\}_{r=1}^m, & u = k, \quad v \in I. \end{cases}$$

Если $n - p > 1$, то идти 3.

Шаг 9. Конец.

Снижение размерности пространства возможно в форме рекурсивного процесса объединения признаков. Набор признаков, полученный на очередном шаге рекурсии, является исходным для алгоритма на следующем шаге. В идеале описание объектов классов может быть сведено к одному латентному признаку. В общем случае завершение рекурсивного процесса объединения признаков определяется условием $\Phi = \emptyset$ при $p = 0$ на 3-м шаге алгоритма.

3. Вычислительный эксперимент

Вычислительный эксперимент основывается на данных, приведенных в работе [6], о 147 пациентах, полученных из Центрального военного госпиталя Министерства обороны Республики Узбекистан. Для описания клинических и функциональных параметров каждого пациента (объекта) было использовано 29 количественных признаков. Диагностировались две группы (класса) пациентов: 111 практически здоровые (K_1) и 36 больные артериальной гипертонией (K_2).

Разбиение на непересекающиеся группы объектов по свойству связанности в рамках конкретной метрики с указанием (для идентификации) их порядковых номеров в каждом классе представлено в табл. 1. В скобках приводится количество объектов, входящих в группу. Число групп является нижней границей для оценки количества

объектов-эталонов покрытия. Подтверждением истинности этого утверждения служат данные из табл. 2, в которой приводится число объектов-эталонов (в скобках) по каждой идентифицированной группе. Общее число объектов-эталонов по взвешенному расстоянию в (4) на базе метрики Чебышева 9, оно больше числа 6 по аналогичному расстоянию на базе метрики Хэмминга.

Мера близости в (4) определяет локальные метрики с весом относительно каждого объекта выборки. Использование локальных метрик позволяет выделять области в признаковом пространстве, которые защищают (притягивают) объекты-эталоны (столпы по терминологии из [2]) покрытия. Число объектов-эталонов покрытия является одним из показателей компактности обучающей выборки, выражаемой через устойчивость логических закономерностей в форме гипершаров. Значение устойчивости гипершара с центром в объекте-эталоне $S \in K_t$, $t = 1, \ldots, l$, вычисляется через мощность множества $D_S = \{S_i \in K_t \mid d(S_i, S) < r_S\}$, где r_S — расстояние до ближайшего от S объекта из CK_t по локальной метрике из (4).

Для доказательства того, что от снижения размерности признакового пространства повышается компактность выборки, было использовано нелинейное отображение значений из наборов (непересекающихся групп) признаков в описании объектов на числовую ось. Алгоритм иерархической агломеративной группировки формирует наборы в порядке, определяемом отношением по значениям (5) и (6). Селекцию признаков можно проводить путем удаления наборов (латентных признаков в новом пространстве) в порядке, обратном их формированию. Зависимость между числом объектов-эталонов покрытия и размерностью пространства при селекции признаков приводится в табл. 3. В скобках указано число исходных признаков, включенных в наборы.

Сравнительный анализ результатов покрытия обучающей выборки по табл. 2 и 3 показывает на значительное уменьшение числа объектов-эталонов при использовании нелинейного отображения определяемых наборов признаков на числовую ось. Число объектов-эталонов покрытия, обеспечивающих корректное распознавание на обучающей выборке, монотонно не возрастает при уменьшении размерности признакового пространства. Удаление шумовых объектов в процессе селекции выборки служит препят-

Т а б л и ц а 1. Число групп объектов по классам

Метрика	Классы	
	Здоровые	Больные
Хемминга	1(111)	2(35), 3(1)
Чебышева	1(111)	2(35), 3(1)

Т а б л и ц а 2. Число объектов-эталонов покрытия в группах

Взвешенное расстояние на базе метрики	Классы	
	Здоровые	Больные
Хемминга	1(3)	2(2), 3(1)
Чебышева	1(5)	2(3), 3(1)

Т а б л и ц а 3. Число объектов-эталонов покрытия при селекции признаков

Взвешенное расстояние на базе метрики	Размерность пространства				
	6(29)	5(28)	4(26)	3(24)	2(17)
Хемминга	5	4	4	2	2
Чебышева	3	3	3	3	2

Т а б л и ц а 4. Разбиение на группы по метрике Журавлева

Класс	Число	
	групп	объектов-эталонов
Здоровые	1	4
Больные	5	9

ствием для переобучения алгоритма распознавания. Как было указано выше, решение этой проблемы связано с обобщающей способностью распознающих алгоритмов и в данной работе не рассматривается.

Вычисление бинарных мер близости со свойствами метрики между объектами в разнотипном признаковом пространстве может производиться с помощью функций, представляющих сумму мер близости по множеству номинальных и количественных признаков. Примерами таких функций служат метрика Журавлева и локальные метрики из работы [5].

Для демонстрации группировки объектов с описанием в разнотипном признаковом пространстве 14 количественных признаков (с 16 по 29) исходной выборки преобразуем в номинальные, используя разбиение их значений на непересекающиеся интервалы по критерию (5). В номинальной шкале градации каждого из 14 перечисленных признаков определяются номерами интервалов, к которым принадлежат их исходные значения. В нашем случае номера для градаций могут выбираться из $\{0, 1\}$ или $\{1, 2\}$. Для сглаживания влияния масштабов измерений на вычисление расстояния между объектами значения количественных признаков с 1 по 15 нормируются в [0..1]. Результаты использования метрики Журавлева на модифицированных данных для разбиения на группы по свойству связанности объектов через оболочки классов приводятся в табл. 4.

Анализ результатов вычислительного эксперимента из табл. 1, 2 и 4 показывает, что число групп разбиения и объектов-эталонов покрытия классов с описанием в разнотипном признаковом пространстве увеличилось по сравнению с аналогичными показателями при описании объектов количественными признаками.

Унификация шкал (путем сведения к единой шкале) измерений для разнотипных признаков позволяет использовать нелинейное отображение их в описании объектов на числовую ось. Достигается это за счет расширения размерности признакового пространства. Наличие или отсутствие значения градации номинального признака в описании объекта отображается в значение из $\{0, 1\}$ и рассматривается как отдельный признак.

Таким образом, показано, что предобработка данных путем комплексного использования методов кластерного анализа для группировки объектов и признаков позволяет значительно сократить объем обучающих выборок и повысить устойчивость логических закономерностей, определяемых на них. В дальнейшем предполагается исследование обобщающей способности решающих правил на основе критериев начала переобучения.

Список литературы / References

[1] **Субботин С.А.** Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов // Математичні машини і системи, 2010. № 1. С. 25–39.

Subbotin, S.A. The complex characteristics and comparison criteria of training sets for diagnostics and pattern recognition // Matematichni mashini i sistemi. 2010. No. 1. P. 25–39. (in Russ.)

[2] **Загоруйко Н.Г., Кутненко О.А., Зырянов А.О., Леванов Д.А.** Обучение распознаванию образов без переобучения // Машинное обучение и анализ данных. 2014. Т. 1, № 7. С. 891–901.
Zagoruiko, N.G., Kutnenko, O.A., Zyryanov, A.O., Levanov, D.A. Learning to recognition without overfitting // Machine Learning and Data Analysis. 2014. Vol. 1, No. 7. P. 891–901. (in Russ.)

[3] **Дюк В.А.** Формирование знаний в системах искусственного интеллекта: геометрический подход // Вестн. Акад. техн. творчества. СПб. 1996. № 2. С. 46–67.
Dyuk, V.A. Creation of knowledge in artificial intelligence systems: geometric approach // Vestnik Akademicheskogo Tekhnicheskogo Tvorchestva. SPb. 1996. No. 2. P. 46–67. (in Russ.)

[4] **Берестнева О.Г., Муратова Е.А., Янковская А.Е.** Анализ структуры многомерных данных методом локальной геометрии // Изв. Томского политехн. ун-та. 2003. Т. 306, № 3. С. 19–23.
Berestneva, O.G., Muratova, E.A., Yankovskaya, A.E. Analysis of the structure of multidimensional data by the local geometry // Bulletin of the Tomsk Polytechnic University. 2003. Vol. 306, No. 3. P. 19–23. (in Russ.)

[5] **Игнатьев Н.А.** Обобщенные оценки и локальные метрики объектов в интеллектуальном анализе данных. Ташкент: Университет, 2014. 72 с.
Ignat'ev, N.A. The generalized estimations and local metrics of objects in data mining. Tashkent: Universitet, 2014. 72 p. (in Russ.)

[6] **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.

Поступила в редакцию 1 апреля 2015 г.,
с доработки — 26 октября 2015 г.

Cluster analysis and choice of standard objects in supervised pattern recognition problems

Ignat'ev, Nikolay A.

National University of Uzbekistan, Tashkent, 100174, Uzbekistan

Corresponding author: Ignat'ev, Nikolay A., e-mail: `n_ignatev@rambler.ru`

Purpose: Search for solutions of the following problems:

— to find the minimum cover of a training set $E_0 = S_1, \ldots, S_m$ by standard objects. The set E_0 is divided into $l (l \geq 2)$ of disjoint subsets (classes) $K_1, \ldots, K_l$. The objects are described by a set of features $X(n) = (x_1, \ldots, x_n)$;

— a reduction of the dimension of the features space by constructing groups from the disjoint sets of features $X(k_1), \ldots, X(k_p)$, $k_1 + \cdots + k_p \leq n$ and nonlinear mapping of their values on the real axis in the description of objects.

Methodology: The method of partitioning the training set E_0 into disjoint groups of objects using the property of their connectivity through the subset of boundary objects (shell) of classes $L(E_0, \rho)$ on a metric $\rho(x, y)$ was developed. The set $G_{t\nu} = S_{\nu_1}, \ldots, S_{\nu_c}$, $c \geq 2$, $G_{t\nu} \subset K_t$, $\nu < |K_t|$ presents an area (group) with the constrained objects in the class K_t, if for any $S_{\nu_i}, S_{\nu_j} \in G_{t\nu}$ the pathway $S_{\nu_i} \leftrightarrow S_{\nu_k} \leftrightarrow \cdots \leftrightarrow S_{\nu_j}$

will exist. The standard objects of minimum covering of training set E_0 are selected for each group $G_{t\nu}$ separately.

The rule of hierarchical clustering of features offers for the nonlinear mapping of their values on the real axis. The use of this rule allows:

— to form a new space from latent features;

— to produce the ordered selection of informative features.

Findings: It is proved, that the number of standard objects for covering which ensures the correct recognition on precedents in the training set, does not increase monotonously when the number of features with low value of information are reduced.

Originality/value: Complex use of methods of cluster analysis for both group objects and features allows to reduce the volume of training sets and to increase the stability of the internally defined logical regularities.

Keywords: pattern recognition, logical regularity, cluster analysis of data, shell of classes, standard objects.

Received 1 April 2015
Received in revised form 26 October 2015