ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2019 Управление, вычислительная техника и информатика № 49

УДК 519.95
DOI: 10.17223/19988605/49/7

Н.А. Игнатьев, А.И. Мирзаев

ОТБОР ПРИЗНАКОВ В СОБСТВЕННОЕ ПРОСТРАНСТВО ОБЪЕКТА НА ОСНОВЕ МЕРЫ ЕГО КОМПАКТНОСТИ

Рассматривается использование логических закономерностей в форме гипершаров для поиска собственного признакового пространства объекта выборки из непересекающихся классов. Разработан алгоритм проверки истинности отношения связанности объектов по системе гипершаров на определяемом наборе признаков. Отношение связанности используется для вычисления значения меры компактности объекта при поиске его собственного признакового пространства.

Ключевые слова: отношение связанности объектов; собственное признаковое пространство; мера компактности.

Понятие «собственное пространство объекта» связанно с принятием решения. Принятие решения зависит от закономерностей (как правило, скрытых), которые наиболее точно передают особенности объекта. В [1] эти особенности было предложено искать в виде логических закономерностей в окрестности объекта. Аргументом в пользу такого подхода служило отсутствие машинных алгоритмов, позволяющих производить поиск логических закономерностей за приемлемое время.

Отбор информативного набора признаков в окрестности объекта и вычисление индекса здоровья (оценки объекта) по нему рассматривался в [2]. В качестве критерия для отбора использовался поиск максимума разности частоты встречаемости представителей двух классов по системе вложенных гипершаров. Центром гипершаров являлся рассматриваемый объект.

Выбор эвристик для построения алгоритмов распознавания образов основывается на гипотезе о компактности объектов классов. Общепринятого определения меры компактности не существует [3]. Показано [4], что в метрических алгоритмах классификации компактность зависит от многообразия структур отношений между объектами классов. Различаются между собой и численные методы для количественного оценивания компактности. В одномерном случае для оценивания используются интервальные методы, в многомерном – вычисление меры компактности объектов классов и выборки в целом по заданной метрике. Общим для одномерного и многомерного случаев является наличие областей признакового пространства, в границах которых вычисляется мера компактности.

В одномерном случае на числовой оси можно производить сравнение объектов по значениям их исходных и латентных признаков, используя отношения «больше», «меньше» или «равно». При вычислении меры компактности в многомерном случае [4] применялось отношение связанности объектов по подмножеству (оболочке) граничных объектов классов по заданной метрике. Связанность объектов S_i, S_j рассматривалось как свойство логических закономерностей в форме гипершаров, центрами которых они являлись. Объекты S_i и S_j считались связанными, если в пересечении их гипершаров были объекты оболочки.

Связанность объектов применялась для анализа кластерной структуры классов с помощью меры компактности. Для вычисления меры компактности в (0; 1] использовались число непересекающихся групп и количество объектов, в них входящих.

Отбор информативных признаков на основе методов кластеризации рассматривался в [5]. Использовалось разбиение признаков на группы, и в каждой группе выделялось по одному наиболее типичному представителю. Результаты группировки существенно зависели от вводимой меры расстояния между признаками.

В [6] в качестве критерия информативности признаков применялась функция конкурентного сходства (FRiS-функция). Среднее значение функции конкурентного сходства зависит от того, как близко группы объектов находятся от разделяющей границы. Те объекты, которые располагаются в тесном окружении своих объектов и значительно удалены от объектов других классов, имеют более высокое значение функции, чем периферийные объекты, близкие к другим классам. Отбор информативных признаков позволяет сделать прозрачным способ построения решающих правил и количественно оценить компактность классов.

При реализации алгоритмов отбора информативных наборов признаков объекта необходимо учитывать:

– наличие или отсутствие свойства инвариантности признаков к масштабам измерений;

– выбор меры близости между объектами со свойствами метрики;

– наличие шумовых объектов в выборке и способов их обнаружения;

– истинность отношения связанности объектов классов;

– выбор способа вычисления меры компактности объекта класса.

В работе определяется бинарное отношение связанности объектов одного отдельно взятого класса обучающей выборки. Это отношение используется для вычисления меры компактности объекта класса с целью отбора признаков в его собственное пространство. Мера компактности рассматривается в качестве индекса объекта по определяемому набору признаков и служит средством для поиска скрытых закономерностей в базах данных.

1. Постановка задачи

Одной из целей анализа кластерной структуры данных в [4] через отношение связанности объектов классов было решение задачи о минимальном покрытии обучающей выборки объектами-эталонами. Объекты каждого класса разбивались на непересекающиеся группы. Отношение связанности объектов гарантировало единственность числа групп и их состава. Поиск объектов–эталонов минимального покрытия производился по каждой группе в отдельности. Среднее число объектов выборки, притягиваемых одним эталоном, использовалось как показатель обобщающей способности алгоритма распознавания. В идеале отдельно взятый объект выборки мог быть единственным эталоном всего класса. Научный и практический интерес представляет оценка вклада объекта в обобщающую способность.

Количественная мера компактности объекта зависит от структуры его отношений с другими объектами обучающей выборки. Среди факторов, влияющих на оценку структуры, особое значение имеет размерность признакового пространства [7] и расстояния между объектами по заданной метрике $\rho(x, y)$. С этими факторами связано такое понятие, как «проклятие размерности пространства» [8].

Задача вычисления меры компактности объекта в рамках его собственного признакового пространства формулируется так. Считается, что задано множество объектов $E_0 = \{S_1, ..., S_m\}$, разделенное на непересекающиеся классы K_1 и K_2. 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 $\rho(x, y)$.

Пусть $r_d = \rho(S_d, S_u) = \min\limits_{S_j \in K_{3-t}} \rho(S_d, S_j)$ – расстояние от $S_d \in K_t$, $t = 1, 2$, до ближайшего (граничного) объекта $S_u \in K_{3-t}$, $\Gamma(K_t, \rho)$ – множество граничных объектов класса K_{3-t}. Обозначим через $O(S_d,\rho) = \{S_i \in K_t | \rho(S_i, S_d) < r_d\}$ и $Z(S_d, \rho) = \{S_i \in O(S_d, \rho) | \rho(S_i, S^*) \le r_d\}$, $S^* \in \Gamma(K_t, \rho)$.

Объекты S_d, $S_u \in K_t$ считаются связанными, если $O(S_u,\rho) \cap Z(S_d,\rho) \neq \varnothing$. Компактность объекта $S_d \in K_t$ на наборе $X(k) \subset X(n)$, $k \le n$, вычисляется как

$$\theta_d(X(k)) = \left|\left\{S_i \in K_t \mid O(S_i, \rho) \cap Z(S_d, \rho) \neq \varnothing\right\}\right| / \left|K_t\right|. \quad (1)$$

Очевидно, что $0 < \theta_d(X(k)) \le 1$, так как $K_1 \cap K_2 = \varnothing$. Требуется найти такой набор $X(u) \subset X(n)$, при котором

$$\theta_d(X(u)) = \max_{X(k) \subset X(n)} \theta_d(X(k)). \quad (2)$$

Определяемый по (2) набор $X(u)$ применяется для описания собственного признакового пространства объекта $S_d \in K_t$, а значение $\theta_d(X(u))$ используется как мера его компактности.

2. Отбор признаков в собственное пространство объекта

При отборе набора $X(k) \subset X(n)$, $k \le n$, для описания признакового пространства объекта $S \in E_0$ необходимо:

– произвести выбор метрики в качестве меры расстояния между объектами;

– задать способ нормирования значений количественных признаков для унификации масштабов измерений;

– определить наличие шумовых объектов в окрестности объекта S.

Описание допустимого объекта в рамках его собственного пространства из информативных признаков необходимо для нахождения индивидуальной меры сходства (различия) с другими объектами. Эта мера должна отражать отношения между объектами и служить средством для принятия решения.

Обозначим через i, j – множество индексов соответственно количественных и номинальных признаков в наборе $X(n)$. Для унификации масштабов измерений значения количественных признаков дробно-линейным преобразованием отобразим в [0; 1]. В качестве меры расстояния между объектами $S_u, S_v \in E_0$ ($S_c = (a_{c1}, \ldots, a_{cn})$, $c = 1, \ldots, m$) будем использовать метрику Журавлёва

$$\rho(S_u, S_v) = \sum_{i \in I} |a_{ui} - a_{vi}| + \sum_{i \in J} \begin{cases} 1, a_{ui} \ne a_{vi}, \\ 0, a_{ui} = a_{vi}. \end{cases}$$

Для выбора информативного набора признаков $X(r) \subset X(n)$, $r \le n$, из собственного пространства объекта предлагается производить предобработку данных. Смысл предобработки заключается в поиске первой пары признаков $(x_i, x_j) \subset X(n)$, $i \ne j$, для информативного набора.

Множество расстояний объектов E_0 от $S_d \in K_t$ по паре (x_i, x_j), $i, j \in \{1, \ldots, n\}$ рассматривается как радиусы вложенных гипершаров, представленных в виде упорядоченной последовательности

$$\rho(S_d, S_{d_1}), \ldots, \rho(S_d, S_{d_\mu}), \ldots, \rho(S_d, S_{d_m}), S_d = S_{d_1}, \quad (3)$$

где $\mu = |K_t|$. Обозначим через $\eta_t(i, j)$($\eta_{3-t}(i, j)$) число ближайших к S_d объектов по (3) из K_t (K_{3-t}) при условии, что $\eta_t(i, j) + \eta_{3-t}(i, j) = \mu$. При $\eta_{3-t}(i, j) = 0$ все объекты класса K_t содержатся в гипершаре с центром в S_d. С помощью значения $\eta_t(i, j)$ решается проблема выбора первого шага при отборе собственного пространства объекта.

В процессе предобработки необходимо исключить появление сходных с $S_d \in K_t$ описаний объектов из K_{3-t}. Плотность распределения представителей класса K_t в окрестности объекта S_d по наборам из $\{(x_i, x_j)\}$ предлагается определять по значениям

$$\theta_{ij} = \max_{1 \le u \le \eta_t(i,j)} \left(z_t(u) - z_{3-t}(u) \right), \gamma_{ij} = \arg \max_{1 \le u \le \eta_t(i,j)} \left(z_t(u) - z_{3-t}(u) \right), \quad (4)$$

где $z_t(u)$($z_{3-t}(u)$) – число объектов класса K_t (K_{3-t}) в гипершаре радиуса $\rho\left(S_d, S_{d_u}\right)$ из последовательности (3).

Пусть $\rho\left(S_d, S_{d_k}\right), k \ge 2$ – значение радиуса гипершара из (3), определяемого по расстоянию до первого ближайшего объекта $S_{d_k} \in K_{3-t}$. Для выбора первой пары признаков в $X(r) \subset X(n)$, $r \le n$, из множества наборов $\{(x_i, x_j)\}$ используется матрица $B(S_d) = \{b_{ij}\}_{n \times n}$, значения элементов которой вычисляются как

$$b_{ij} = \begin{cases} 0, \rho(S_d, S_{d_k}) = 0, \\ |O(S_d, \rho)| \times \theta_{ij} / \gamma_{ij}, \rho(S_d, S_{d_k}) > 0. \end{cases} \quad (5)$$

Целью вычисления значений $b_{ij} \ne 0$ является поиск кластеров данных с максимальной плотностью объектов одного с $S_d \in K_t$ класса.

Существует зависимость количества связанных с $S_d \in K_t$ объектов для вычисления (2) от наличия или отсутствия шумовых объектов. Шумовые объекты из $K_{3\text{-}t}$ по $X(p)$, $p \le n$, и метрике $\rho(x, y)$ для объекта $S_d \in K_t$ предлагается определять следующим образом.

Пусть $\Gamma(p)$ – множество граничных объектов классов по $X(p) \subset X(n)$, $p \le n$, и метрике $\rho(x, y)$, $G(p) = \{g_i\}_{S_i \in \Gamma(p)}$, где $g_i = \left| \left\{ S_j \in K_{3-c} \middle| \rho(S_j, S_i) = \min_{S_r \in K_c} \rho(S_j, S_r) \right\} \right|$ – число объектов, для которых $S_i \in K_c \cap \Gamma(p)$, $c = 1, 2$, является граничным. Объект $S_i \in K_{3\text{-}t} \cap \Gamma(p)$, $t = 1, 2$, считается шумовым относительно объекта $S_d \in K_t$ если:

1) $\rho(S_d, S_i) = \min_{S_r \in K_{3-t}} \rho(S_d, S_r)$;

2) $g_i/|K_t| > |O(S_i, \rho)|/|K_{3\text{-}t}|$.

Результаты предобработки с использованием (3), (4) в виде пары признаков $H(2) = (x_i, x_j)$ рассматриваются в качестве начального приближения эвристического алгоритма пошагового отбора информативных признаков объекта $S_d \in K_t$. Последовательность шагов по реализации алгоритма такова.

Шаг 1. Ввод $H(2)$. $p = 2$. $X(p) = H(2)$. $count = 0$. $T = 0$.

Шаг 2. По набору $X(p)$ вычислить множество граничных объектов $\Gamma(p)$ и $G(p) = \{g_i\}_{S_i \in \Gamma(p)}$.

Шаг 3. Определить объект $S_c \in K_{3\text{-}t} \cap \Gamma(p)$ с $\rho(S_c, S_d) = \min_{S_r \in K_{3-t} \cap \Gamma(p)} \rho(S_d, S_r)$. Если $g_c/|K_t| > O(S_c, \rho)/|K_{3\text{-}t}|$, то обновить множество граничных объектов $\Gamma(p)$ по $X(p)$ на $E_0 \backslash \{S_c\}$.

Шаг 4. Вычислить значения элементов множества $Z(S_d, \rho)$ и $\theta_d(X(p))$ по (1). Если $\theta_d(X(p)) > T$, то $T = \theta_d(X(p))$, $H(p) = X(p)$, $u = p$. Если $T = 1$, то идти 7.

Шаг 5. $count = count + 1$. Если $count = n$, то идти 7.

Шаг 6. $R = 0$. $v = 0$. **Начало цикла:** Для всех $x_a \in X(n) \backslash X(p)$ **вычислять** значения θ_{ij}, γ_{ij} по (3) и (4) на $X(p + 1) = X(p) \cup \{x_a\}$. Если $|O(S_d, \rho)| \times \theta_{ij}/\gamma_{ij} > R$, то $R = |O(S_d, \rho)| \times \theta_{ij}/\gamma_{ij}$, $v = a$. **Конец цикла.** $X(p + 1) = X(p) \cup \{x_v\}$. $p = p + 1$. Идти 2.

Шаг 7. Вывод T, $H(u)$.

Шаг 8. Конец.

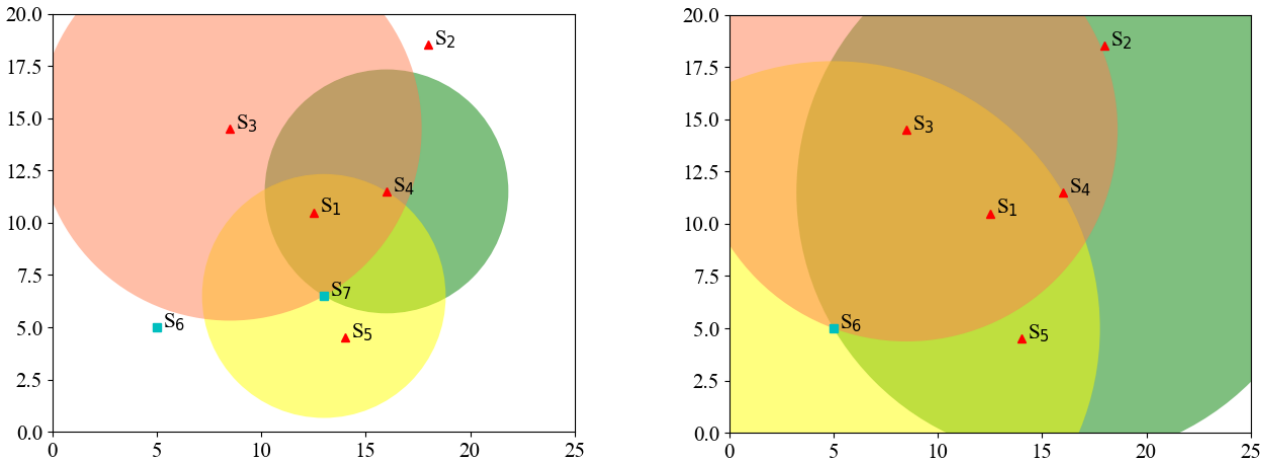

Рис. 1. Иллюстрация процесса выбора связанных объектов для S_4 до и после удаления шумового объекта S_7
Fig. 1. Illustration of the process of selecting connected objects for S_4 before and after removing the S_7 noise object

Процесс выбора связанных объектов для вычисления меры компактности (1) до и после удаления шумового объекта показан на рис. 1.

3. Вычислительный эксперимент

Для эксперимента была взята выборка данных German из [9]. Выборка представлена 1 000 объектами, разделенными на два класса K_1 и K_2. Каждый объект рассматривается как кредитная история

клиента банка. Кредитная история описывается 20 признаками, 7 из которых измеряются в количественных шкалах, 13 – в номинальных. В табл. 1 и 2 представлена последовательность отбора признаков в собственное пространство для объектов $S_{907} \in K_1$ и $S_5 \in K_2$ алгоритмом из разд. 2. Количество объектов в гипершаре и связанных получено после удаления шумовых объектов.

Таблица 1

Отбор признаков в собственное пространство объекта $S_{907} \in K_1$

Набор признаков	Количество объектов		Значение (1)
	в гипершаре	связанных	
x_9, x_{20}	23	23	0,0000
x_5, x_9, x_{20}	32	31	0,0443
x_5, x_6, x_9, x_{20}	63	58	0,0829
x_5, x_6, x_9, x_{14}, x_{20}	55	54	0,0771
x_5, x_6, x_9, x_{12}, x_{14}, x_{20}	59	90	0,1286
x_5, x_6, x_9, x_{12}, x_{14}, x_{16}, x_{20}	49	85	0,1214
x_5, x_6, x_9, x_{11}, x_{12}, x_{14}, x_{16}, x_{20}	77	96	0,1371
x_2, x_5, x_6, x_9, x_{11}, x_{12}, x_{14}, x_{16}, x_{20}	72	91	0,1300
x_2, x_5, x_6, x_9, x_{11}, x_{12}, x_{14}, x_{15}, x_{16}, x_{20}	63	101	0,1443

На наборе x_9, x_{20} (см. табл. 1) значение (1) равно 0, так как существует объект из K_1, описание которого совпадает (пересекается) с описанием ближайшего к $S_{907} \in K_1$ объекта из K_2.

Таблица 2

Отбор признаков в собственное пространство объекта $S_8 \in K_2$

Набор признаков	Количество объектов		Значение (1)
	в гипершаре	связанных	
x_8, x_{13}	7	16	0,0000
x_4, x_8, x_{13}	16	23	0,0000
x_4, x_8, x_{13}, x_{14}	16	24	0,0000
x_4, x_8, x_{13}, x_{14}, x_{18}	18	19	0,0000
x_4, x_8, x_{11}, x_{13}, x_{14}, x_{18}	37	41	0,0586
x_4, x_5, x_8, x_{11}, x_{13}, x_{14}, x_{18}	42	48	0,0686

Как видно из табл. 1 и 2 количество связанных объектов может быть больше, меньше или равно количеству объектов в гипершаре. Собственные наборы признаков для восьми случайно выбранных объектов из классов K_1 и K_2 приведены в табл. 3.

Таблица 3

Результаты отбора признаков в собственное пространство восьми объектов из классов K_1 и K_2

Номер объекта (класс)	Набор признаков	Значение (1)
310 (1)	x_5, x_{10}, x_{11}, x_{12}, x_{14}, x_{18}, x_{20}	0,0400
325 (1)	x_1, x_5, x_{12}	0,0871
460 (1)	x_1, x_2, x_5, x_7, x_8, x_9, x_{10}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{18}, x_{19}, x_{20}	0,1171
826 (1)	x_2, x_3, x_4, x_5, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}	0,0771
38 (2)	x_2, x_4, x_5, x_6, x_7, x_{16}, x_{17}	0,0167
125 (2)	x_1, x_5, x_{11}, x_{19}	0,0167
707 (2)	x_2, x_3, x_5, x_6, x_{10}, x_{12}, x_{14}, x_{15}, x_{16}, x_{18}, x_{19}, x_{20}	0,0300
827 (2)	x_1, x_2, x_5, x_6, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{17}, x_{18}, x_{20}	0,0267

Как видно из табл. 3, свойством собственного пространства объектов класса K_2 (плохие клиенты) является относительно низкое значение компактности по отношению связанности объектов (см. значение (1)) по сравнению с K_1 (хорошие клиенты).

Для сравнительного анализа рассмотрим формирование собственного признакового пространства объектов из табл. 3 по критерию из [10]. Устойчивость объекта $S_d \in K_t$ по (3) на наборе $X(k) \subset X(n)$ вычисляется как значение функционала

$$F\left(S_d, X\left(k\right)\right) = \max_{1 \le i \le m} \left(\frac{z_t\left(i\right)}{\left|K_t\right|} - \frac{z_{3-t}\left(i\right)}{\left|K_{3-t}\right|} \right), \qquad (6)$$

где $z_t(i)$, $z_{3-t}(i)$ – число объектов в $\{S_{d_1},...,S_{d_i}\} \subset E_0$, определяемых по (3), соответственно из классов K_t и K_{3-t}. Множество допустимых значений (6) принадлежит интервалу (0; 1]. Условием для поиска набора информативных признаков $X(\mu)$, $\mu \le n$, для $S_d \in K_t$ является

$$F\left(S_d, X\left(\mu\right)\right) = \max_{1 \le k \le n} \max_{\{X(k)\}} F\left(S_d, X\left(k\right)\right). \quad (7)$$

Выбор первой пары признаков в $X(2)$ из $\{(x_i, x_j)\}_{i,j \in \{1,\ldots,n\}}$ производится по (7). Процесс отбора реализован в виде последовательного (пошагового) добавления признаков в $X(\mu)$, μ = 2, 3,… Существенным отличием отбора информативных признаков для $S_d \in K_t$ по (7) от (1) является «безразличие» к наличию шумовых объектов из класса K_{3-t}. Информативные наборы признаков, получаемые по экстремуму (7), приведены в табл. 4.

Таблица 4

Информативные наборы признаков объектов по критерию (7)

Номер объекта (класс)	Набор признаков	Значение (7)
310 (1)	x_2, x_5, x_{14}	0,2238
325 (1)	x_1, x_3, x_5	0,3819
460 (1)	x_1, x_2, x_5, x_{14}, x_{16}	0,3571
826 (1)	x_3, x_5, x_9, x_{19}	0,2481
38 (2)	x_6, x_{10}, x_{13}, x_{16}, x_{20}	0,2090
125 (2)	x_1, x_2, x_7, x_{15}, x_{20}	0,2143
707 (2)	x_2, x_4, x_{20}	0,2338
827 (2)	x_1, x_{13}, x_{14}, x_{16}, x_{20}	0,3014

Относительно малое различие между оценками по (7) из разных классов (см. табл. 4) объясняется отсутствием учета наличия шумовых объектов и непустого множества объектов из K_1 и K_2, описания которых на определяемом наборе признаков $X(u) \subset X(n)$ совпадают.

Заключение

Разработан метод отбора признаков в собственное пространство объекта на основе отношения связанности объектов по системе гипершаров. Отношение связанности применяется для вычисления меры компактности объекта относительно своего класса в определяемом подпространстве заданного признакового пространства. Метод может быть использован для поиска скрытых закономерностей в данных в рамках информационных моделей из слабо формализованных предметных областей.

ЛИТЕРАТУРА

1. Дюк В.А. Методология поиска логических закономерностей в предметной области с нечеткой системологией: на примере клинико-экспериментальных исследований : дис. ... д-ра техн. наук. СПб., 2005. 309 с.
2. Ignat'ev N.A., Mirzaev A.I. The Intelligent Health Index Calculation System // Pattern Recognition and Image Analysis. 2016. V. 26, No. 1. P. 73–77.
3. Загоруйко Н.Г., Борисова И.А., Дюбанов В.В., Кутненко О.А. Количественная мера компактности и сходства в конкурентном пространстве // Сибирский журнал индустриальной математики. 2010. Т. 13, № 1 (41). С. 59–71.
4. Ignatyev N.A. Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis. 2018. V. 28, No. 4. P. 590–597.
5. Колесникова С.И. Методы анализа информативности разнотипных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2009. № 1 (6). С. 69–80.
6. Загоруйко Н.Г., Кутненко О.А., Борисова И.А., Дюбанов В.В., Леванов Д.А., Зырянов О.А. Выбор информативных признаков для диагностики заболеваний по генетическим данным // Вавиловский журнал генетики и селекции. 2014. Т. 18, No. 4/2. С. 898–903.
7. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. Мо. : Финансы и статистика, 1989. 607 с.
8. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение : пер. с англ. М. : ДМК Пресс, 2018. 652 с.
9. The UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml/datasets (accessed: 09.04.2019).
10. Игнатьев Н.А. Индексирование объектов по индивидуальным наборам информативных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2016. № 4 (37). С. 27–35.

Поступила в редакцию 17 апреля 2019 г.

Ignatyev A.N., Mirzaev A.I. (2019) SELECTION OF FEATURES INTO THE OBJECT'S OWN SPACE BASED ON THE MEASURE OF ITS COMPACTNESS. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science]. 49. pp. 55–62

DOI: 10.17223/19988605/49/7

Considered using of logical regularities in the form of hyper-sphere for the selection of feature into the own space of the object from sample $E_0 = \{S_1,\ldots,S_m\}$. Sample divided into of disjoint subset (classes) K_1 and K_2. The objects are described by means of the set of n diverse features $X(n) = (x_1, \ldots, x_n)$. On the set E_0, the metric $\rho(x, y)$ is given.

Let $r_d = \rho(S_d, S_u) = \min\limits_{S_j \in K_{3-t}} \rho(S_d, S_j)$ be the distance from $S_d \in K_t$, $t = 1, 2$ to the nearest (boundary) object $S_u \in K_{3-t}$, $\Gamma(K_t, \rho)$ be the set of boundary objects of the class K_{3-t}. Denote by $O(S_d, \rho) = \{S_i \in K_t | \rho(S_i,S_d) < r_d\}$ and $Z(S_d, \rho) = \{S_i \in O(S_d, \rho) | \rho(S_i, S^*) \le r_d\}$, $S^* \in \Gamma(K_t, \rho)$. The objects $S_d, S_u \in K_t$ are considered to be connected if $O(S_u, \rho) \cap Z(S_d, \rho) \neq \varnothing$. The compactness of the object $S_d \in K_t$ on the set $X(k) \subset X(n)$, $k \le n$ is calculated as

$$\theta_d(X(k)) = \left|\left\{S_i \in K_t \mid O(S_i,\rho) \cap Z(S_d,\rho) \neq \varnothing\right\}\right| / \left|K_t\right|.$$

The object $S_i \in K_{3-t} \cap \Gamma(p)$, $t = 1, 2$ is called noise relative to the object $S_d \in K_t$ if:

1. $\rho(S_d, S_i) = \min\limits_{S_r \in K_{3-t}} \rho(S_d, S_r)$;

2. $g_i/|K_t| > |O(S_i,\rho)|/|K_{3-t}|$, $g_i = \left|\left\{S_j \in K_t \left| \rho(S_j, S_i) = \min\limits_{S_r \in K_{3-t}} \rho(S_j, S_r)\right.\right\}\right|$ is the number of objects, where $S_i \in K_{3-t} \cap \Gamma(p)$ is the nearest one.

The set $X(u) \subset X(n)$, computed on $E_0 \backslash \{S_i\}$ as $\theta_d(X(u)) = \max\limits_{X(k) \subset X(n)} \theta_d(X(k))$ is considered informative for the object $S_d \in K_t$, and the value of $\theta_d(X(u))$ is considered as a measure of its compactness.

To implement the algorithm of step by step selection of features into an informative set, data preprocessing is performed. The purpose of preprocessing is to select the first pair (x_i, x_j) into an informative set based on the proposed criterion. The criterion is used to search for a cluster of data with a maximum density of descriptions of objects of one with S_d class K_t by sets of $\{(x_i, x_j)\}$.

The results of the computational experiment are described according to 1 000 bank customers. Customers are divided into 700 good and 300 bad customers. From the results of the experiment, it was concluded that the measure of compactness among good customers is higher than that of bad ones.

Keywords: relation of connectedness of objects; object's own space; measure of compactness.

IGNATEV Nikolay Aleksandrovich (Doctor of Physics and Mathematics, Professor, National University of Uzbekistan, Tashkent, Uzbekistan).
E-mail: ignatev@rambler.ru

MIRZAEV Aziz Ibrakhimovich (National University of Uzbekistan, Tashkent, Uzbekistan).
E-mail: mirzaevaziz@gmail.com

REFERENCES

1. Dyuk, V.A. (2005) *Metodologiya poiska logicheskikh zakonomernostey v predmetnoy oblasti s nechetkoy sistemologiey: na primere kliniko-eksperimental'nykh issledovaniy* [Methodology of the search for logical patterns in the subject area with fuzzy systemology: clinical and experimental studies]. Engineering Dr. Diss. St. Petersburg.
2. Ignatiev, N.A. & Mirzaev, A.I. (2016) The Intelligent Health Index Calculation System. *Pattern Recognition and Image Analysis*. 26(1). pp. 73–77. DOI: 10.1134/S1054661816010089
3. Zagoruiko, N.G., Borisova, I.A., Dyubanov, V.V. & Kutnenko, O.A. (2010) A quantitative measure of compactness and similarity in the competitive space. *Sibirskiy zhurnal industrial'noy matematiki – Siberian Journal of Industrial Mathematics*. 1(41). pp. 59–71.
4. Ignatyev, N.A. (2018) Structure Choice for Relations between Objects in Metric Classification Algorithms. *Pattern Recognition and Image Analysis*. 28(4). pp. 590–597. DOI: 10.1134/S1054661818040132
5. Kolesnikova, S.I. (2009) Methods for analyzing the informativeness of various types of features. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 1(6). pp. 69–80.
6. Zagoruiko, N.G., Kutnenko, O.A., Borisova, I.A., Dyubanov, V.V., Levanov, D.A. & Zyryanov, O.A. (2014) Feature selection in for medical diagnostics on microarray data. *Vavilovskiy zhurnal genetiki i selektsii – Vavilov Journal of Genetics and Breeding*. 18(4/2). pp. 898–903.

7. Ayvazyan, S.A., Buchstaber, V.M., Enyukov, I.S. & Meshalkin, L.D. (1989) *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti* [Application statistics. Classification and reduction of dimension]. Moscow: Finansy i statistika.
8. Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep Learning*. Cambridge: MIT Press.
9. UCI. (n.d.) [Online] Available from: http://archive.ics.uci.edu/ml/datasets. (Accessed: 9th April 2019).
10. Ignatiev, N.A. (2016) Indexation of objects according to individual sets of informative feature. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 4(37). pp. 27–35. DOI: 10.17223/19988605/37/3