ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2016 Управление, вычислительная техника и информатика № 4 (37)

УДК 519.95
DOI: 10.17223/19988605/37/3

Н.А. Игнатьев

ИНДЕКСИРОВАНИЕ ОБЪЕКТОВ ПО ИНДИВИДУАЛЬНЫМ НАБОРАМ ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Рассматриваются критерии для оценивания (индексирования) объекта в задачах распознавания с учителем. Значение оценки вычисляется как экстремум критерия по индивидуальному набору информативных признаков объекта. Проверяется истинность гипотезы, что в окрестности каждого объекта существует устойчивая логическая закономерность.

Ключевые слова: распознавание образов; индекс объекта; логические закономерности; информативные признаки объекта.

Потребность в индексировании объектов возникает при принятии решений в различных предметных областях. Значения индексов востребованы для мониторинга рынка купли-продажи ценных бумаг, экологического состояния окружающей среды, уровня террористической угрозы, оценки степени социального благополучия общества, цитируемости научных публикаций и т.д. Вычисление значений, как правило, производится по строго фиксированным наборам показателей.

Информативные признаки (показатели), определяемые для всей обучающей выборки, не отражают специфику закономерностей, присущих той или иной области признакового пространства. В [1] утверждается, что для каждого объекта существует своя логическая закономерность, для обнаружения которой предлагалось использовать локальные метрики. Применение локальных метрик основано на эвристиках, так как чётких критериев их выбора не существует. Интерес представляет разработка методов отбора информативных признаков инвариантных к масштабам измерений данных, комбинаторная сложность реализации которых позволяет получать результаты за приемлемое время.

Самую простую и легко интерпретируемую структуру, задаваемую отношениями на элементах непустого множества, представляет линейный порядок. Как правило, первыми кандидатами для включения в информативный набор являются независимые признаки. Примером использования линейного порядка является отбор наборов признаков с максимально выраженной независимостью, который применялся при синтезе искусственных нейронных сетей с минимальной конфигурацией в [2].

Потребность в использовании индивидуального набора информативных признаков объекта для принятия решений возникает при постановке диагноза болезни, разработке мер по предотвращению техногенных катастроф на конкретной территории. В медицинской практике при одном и том же диагнозе у двух человек причинами (диагностическими признаками) болезни могут быть разные симптомы и синдромы.

Метод отбора индивидуальных информативных наборов признаков с помощью локальных метрик объектов описан в [3]. Для отбора использовался критерий на основе максимальной разницы частот встречаемости представителей (объектов) двух классов K_1 и K_2 в последовательности, упорядоченной по локальной метрике объекта. Экстремальное значение критерия, полученное по медицинским данным с описанием состояния больных и практически здоровых индивидуумов, предлагалось интерпретировать как индекс здоровья.

Индивидуальный набор информативных признаков допустимого объекта позволяет:

– выделять логические закономерности в его окрестности;

– объяснять процесс принятия решения при распознавании;

– определять принадлежность к шумовым (аномальным) объектам классов;

– производить выбор опорных множеств признаков в моделях алгоритмов распознавания.

В данной работе для отбора индивидуальных наборов информативных признаков предложено два новых критерия, отличных от описанного в [3]. Также как и в [3], при вычислении значений по этим критериям используются функции близости по определяемым наборам признаков. Упорядочение объектов по значениям функции близости в зависимости от поставленных целей позволяет определять:

– устойчивость логической закономерности относительно исследуемого объекта;

– границу между представителями двух классов, степень истинности гипотезы о компактности при которой максимальна.

Свойство инвариантности к масштабам шкал измерений является атрибутом критерия нелинейного отображения групп исходных (сырых) разнотипных (номинальных и количественных) признаков на числовую ось. По аналогии с методом локальной геометрии [1] начало координат размещается в исследуемом объекте. При нелинейном отображении наряду с синтезом латентных признаков по критерию группировки происходит их упорядочение по степени информативности. Существует вывод аналитического представления (формул) для вычисления значений латентных признаков из исходных [4].

1. Критерии отбора индивидуальных наборов информативных признаков

Рассматривается задача распознавания в стандартной постановке. Объекты обучения заданы через множество $E_0 = \{S_1,\ldots,S_m\}$, разделённое на два непересекающихся подмножества (класса) K_1 и K_2, $E_0 = K_1 \cup K_2$. 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.

Обозначим через I, J множество индексов соответственно количественных и номинальных признаков. Считается, что заданы критерии для отбора информативных признаков объекта $S \in E_0$. Требуется по каждому критерию для указанного объекта $S \in E_0$ определить:

– информативный набор признаков $X(k) = \{x_i\}_{i \in I \cup J}$, $k \geq 1$;

– оценку объекта S как экстремальное значение критерия на информативном наборе $X(k)$.

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

Для унификации масштабов измерений значения количественных признаков дробно-линейным преобразованием отображаются в [0,1]. В качестве меры близости между объектами $S_a = (x_{a1},\ldots,x_{an})$ и $S_b = (x_{b1},\ldots,x_{bn})$ используется метрика Журавлёва

$$\rho(S_a, S_b) = \sum_{i \in I} |x_{ai} - x_{bi}| + \sum_{i \in J} \begin{cases} 1, x_{ai} \neq x_{bi}, \\ 0, x_{ai} = x_{bi}. \end{cases}$$

Положим, что для объекта $S_d \in K_p$ по набору признаков $X(k)$, $k \leq n$, построена

$$S_{d_0},\ldots,S_{d_{m-1}},S_{d_0} = S_d, \tag{1}$$

упорядоченная последовательность объектов E_0, отношения между которыми определяются неравенствами вида $\rho(S_{d_i}, S_d) \leq \rho(S_{d_{i+1}}, S_d)$. Для оценки объекта $S_d \in K_p$ по (1) используется функционал

$$F(S_d, X(k)) = \max_{0 \leq i \leq m-1} \left(\frac{z_p(i)}{|K_p \cap E_0|} - \frac{z_{3-p}(i)}{|K_{3-p} \cap E_0|} \right), \tag{2}$$

где $z_p(i), z_{3-p}(i)$ – число объектов в $\{S_{d_0},\ldots,S_{d_i}\} \subset E_0$, определяемых по (1) соответственно из класса K_p и K_{3-p}. Множество допустимых значений (2) принадлежит интервалу (0,1].

Как отдельную задачу можно рассматривать отбор для $S_d \in K_p$ набора информативных признаков $X(\mu)$, $\mu \leq n$, на котором

$$F(S_d, X(\mu)) = \max_{1 \leq k \leq n} \max_{\{X(k)\}} F(S_d, X(k)). \tag{3}$$

Значение (3) для допустимого объекта S_d по набору медицинских показателей $X(\mu)$ в [3] интерпретировалось как индекс здоровья по классу K_p, $p = 1,2$.

В [3] для нахождения экстремума (3) использовались эвристические пошаговые алгоритмы отбора. Было показано, что различные схемы отбора (последовательным удалением малоинформативных либо последовательным включением наиболее информативных) признаков не давали схожих результатов.

Существенным препятствием для эффективного использования (3) в эвристических алгоритмах отбора является большая размерность признакового пространства. Значения близости между объектами становятся размытыми, экспоненциально растёт сложность вычислений. Для уменьшения комбинаторной сложности вычислений предлагается применять предобработку данных.

Идея использования порядка следования разнотипных признаков по степени их независимости для синтеза моделей искусственных нейронных сетей с минимальной конфигурацией описана в [2]. С этой целью формировалась матрица парных близостей (различий) между признаками. Для унификации шкал измерений использовалось преобразование количественных признаков в номинальные по специальному критерию. Порядок следования признаков определялся по матрице парных близостей (различий).

В отличие от [2] в данной работе предлагается использовать значения матрицы близости для пар признаков $(x_i, x_j) \subset X(n)$ по метрике Журавлёва без унификации шкал (сведения к одной шкале) измерений. Элементы матрицы близости $B(S) = \{b_{ij}\}_{n\times n}$ объекта $S \in K_t$ вычисляются как

$$b_{ij} = \begin{cases} \dfrac{1}{2|K_{3-t}|}\sum\limits_{S_u\in K_{3-t}} \rho(S,S_u) - \dfrac{1}{2|K_t|-1}\sum\limits_{S_u\in K_t}\rho(S,S_u), x_i, x_j \in X(n), i \neq j, \\ 0, i = j. \end{cases} \quad (4)$$

Очевидно, что $|b_{ij}| \le 1$.

Обозначим через P множество, значениями элементов которого являются номера исходных признаков. Для выбора по элементам (4) матрицы $B(S) = \{b_{ij}\}_{n\times n}$ упорядоченного набора $X(k) = (x_1,\ldots,x_k)$, $2 \le k \le n$, используется рекурсивная процедура построения последовательности признаков

$$x_{S_1}, x_{S_2}, \ldots, x_{S_n}. \quad (5)$$

Положим $P = \varnothing$. Из матрицы $B(S)$ выделяется пара (x_i, x_j) с наибольшим значением b_{ij} и включается (слева направо) в (5). Номера выделенных признаков фиксируются в $P = P \cup \{i, j\}$. Порядок следования в (x_i, x_j) выбирается из условия $\max\limits_{\mu\in P} b_{i\mu} \ge \max\limits_{\mu\in P} b_{j\mu}$. Каждая следующая пара признаков из $\{1,\ldots,n\}\backslash P$ для (5) по аналогичному принципу определяется из $B(S)$ после удаления в ней строк и столбцов с номерами i и j.

С целью сокращения комбинаторной сложности алгоритмов из (5) удаляется (справа налево) определяемое число r $(0 < r < n)$ элементов. Набор $x_{S_1},\ldots,x_{S_{n-r}}$ является исходным для начала процесса отбора информативных признаков по (3).

Предлагается ещё один, отличный от (3), способ использования последовательности (1) для отбора информативных признаков. Пусть u_i^1, u_i^2 – количество объектов класса K_i, i = 1,2, соответственно в интервалах $[c_1,c_2],(c_2,c_3]$, η – порядковый номер из последовательности (1), $c_1 = 0$, $c_2 = \rho(S_d, S_{d_\eta})$, $c_3 = \rho(S_d, S_{d_{m-1}})$.

Критерий для определения границы c_2 основывается на проверке гипотезы (утверждения) о том, что каждый из двух интервалов $[c_1,c_2]$, $(c_2,c_3]$ содержит значения расстояния $\rho(x, y)$ до объектов только одного класса. Экстремальное значение критерия на наборе $X(k)$, $2 \le k \le n$, вычисляется как

$$R(S_d, X(k)) = \left(\frac{\sum\limits_{i=1}^{2} u_i^1(u_i^1-1) + u_i^2(u_i^2-1)}{\sum\limits_{i=1}^{2}|K_i|(|K_i|-1)}\right)\left(\frac{\sum\limits_{d=1}^{2}\sum\limits_{i=1}^{2} u_i^d(|K_{3-i}| - u_{3-i}^d)}{2|K_1||K_2|}\right) \to \max_{c_1<c_2<c_3}, \quad (6)$$

а множество его допустимых значений принадлежит (0,1]. Выражение в левых скобках (6) представляет внутриклассовое сходство, в правых – межклассовое различие. Информативный набор признаков $X(\mu)$ по (6) определяется как

$$R\left(S_d, X(\mu)\right) = \max_{\{X(k)\}} R\left(S_d, X(k)\right).$$

Обозначим $\lambda_1(t) = |\{S_a \in K_t \mid \rho(S_d,S_a) \in [c_1,c_2]\}|$, $\lambda_2(t) = |\{S_a \in K_{3-t} \mid \rho(S_d,S_a) \in [c_1,c_2]\}|$, $\theta_1(t) = \lambda_1(t)/|K_t|$, $\theta_2(t) = \lambda_2(t)/|K_{3-t}|$, где интервал $[c_1,c_2]$ получен по набору признаков $X(k)$ по (6). Оценка (устойчивость) объекта $S_d \in K_t$ по набору $X(k)$ вычисляется как $U(S_d, X(k)) = \theta_1(t)(1 - \theta_2(t))$ и

$$U\left(S_d, X(\mu)\right) = \max_{\{X(k)\}} U\left(S_d, X(k)\right), \tag{7}$$

где $X(\mu)$ – информативный набор признаков. Вычисление значения оценки по (7) характеризует вид критерия как мультипликативный, а аналогично (3) – как аддитивный.

Логическая закономерность в форме гипершара с центром $S_d \in K_p$ по набору $X(k)$ определяется множеством

$$\varphi(S_d, X(k)) = \{S_a \in K_p \mid \rho(S_d,S_a) < \rho(S_d,S_b)\},$$

где $S_b \in K_{3-p}$ – ближайший к S_d объект из противоположного класса. Для выбора и интерпретации информативного набора признаков объекта S_d и его оценки предлагается использовать экстремум критерия устойчивости логической закономерности в форме гипершара

$$F(S_d, X(\mu)) = \max_{\{X(k)\}} \frac{\left|\varphi\left(S_d, X(k)\right)\right|}{\left|K_p\right|}. \tag{8}$$

2. Выбор латентных признаков для описания объекта

Рассматривается стандартная постановка задачи распознавания, аналогичная описанной в п. 1. Производится выбор собственного признакового пространства объекта $S_d \in E_0$, $d = 1,\ldots,m$, с помощью алгоритма иерархической агломеративной группировки [4]. Алгоритм группировки разбивает набор признаков $X(n)$ на непересекающиеся группы $X(k_1),\ldots, X(k_\tau)$, $k_1+\ldots+ k_\tau \le n$. Нелинейное отображение представителей каждой группы на числовую ось образует новый латентный признак в описании объекта. Считается, что задан критерий для проверки истинности гипотезы о компактности по значениям латентного признака через произведение внутриклассового сходства и межклассового различия. Требуется определить признак с максимальным значением критерия.

Для выбора латентных признаков в собственном пространстве объекта $S_d \in E_0$, $S_d = (a_{d1},\ldots,a_{dn})$, произведём предобработку данных следующим образом. Значения признаков объекта $S = (b_1,\ldots,b_n)$, $S \in E_0$, преобразуем как

$$b_i = \begin{cases} \left|a_{di} - b_i\right|, i \in I, \\ 1, a_{di} = b_i, i \in J, \\ 0, a_{di} \ne b_i, i \in J. \end{cases} \tag{9}$$

Преобразованные по (9) признаки считаются измеренными в количественной шкале измерений, множество номеров которых идентифицируются как $I = \{1,\ldots,n\}$. Для вычисления значений латентных признаков используются правила иерархической агломеративной группировки. Латентные признаки, полученные на p-м шаге группировки, обозначаются как x_j^p, $j \in I$, $p \ge 0$. При $p = 0$, $|I| = n$. Упорядоченное множество значений признака x_j^p объектов из E_0 аналогично (6) разделим на два интервала $[c_1^{jp}, c_2^{jp}]$, $(c_2^{jp}, c_3^{jp}]$, каждый из которых рассматривается как градация номинального признака.

Пусть u_i^1, u_i^2 – количество значений признака x_j^p, $j \in I$, класса K_i, $i = 1, 2$, соответственно в интервалах $[c_1^{jp}, c_2^{jp}], (c_2^{jp}, c_3^{jp}]$, $|K_i| > 1$; 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 $c_1^{jp} = r_{j_1}, c_2^{jp} = r_{j_v}, c_3^{jp} = r_{j_m}$. Аналогичный (6) критерий

$$\left(\frac{\sum_{i=1}^{2} u_i^1\left(u_i^1-1\right)+u_i^2\left(u_i^2-1\right)}{\sum_{i=1}^{2}\left|K_i\right|\left(\left|K_i\right|-1\right)}\right)\left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2} u_i^d\left(\left|K_{3-i}\right|-u_{3-i}^d\right)}{2\left|K_1\right|\left\|K_2\right|}\right)\to \max_{c_1^{jp}<c_2^{jp}<c_3^{jp}} \tag{10}$$

позволяет вычислять оптимальное значение границы c_2^{jp} для интервалов $[c_1^{jp}, c_2^{jp}]$ и $(c_2^{jp}, c_3^{jp}]$.

Экстремум критерия (10) используется в качестве веса w_j^p $(0\le w_j^p \le 1)$ признака x_j^p. При $w_j^p=1$ значения признака x_j^p у объектов из классов K_1 и K_2 не пересекаются между собой.

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

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

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

где w_i^p, w_j^p, w_{ij}^p – веса признаков, определяемые по (10) соответственно по множеству значений x_i^p, x_j^p и их произведения $x_i^p\ x_j^p$ на E_0, значения $t_{ij}, t_i, t_j\in\{-1,1\}$, $\eta_{ij}\in[0,1]$ выбираются по экстремуму функционала

$$\varphi(p,i,j)=\frac{\min\limits_{S_r\in K_1} b_{rij}^p-\max\limits_{S_r\in K_2} b_{rij}^p}{\max\limits_{S_r\in E_0} b_{rij}^p-\min\limits_{S_r\in E_0} b_{rij}^p}=\max_{t_{ij}, t_i, t_j\in\{-1,1\},\ \eta_{ij}\in[0,1]}. \tag{11}$$

Экстремум функционала (11) интерпретируется как отступ между объектами классов K_1 и K_2 по множеству значений по паре признаков (x_i^p, x_j^p), $0\le p<n$, $i, j\in I$, $i\ne j$.

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

$$z_{ij}^p=\begin{cases} w_i^p, i=j, \\ \text{значению (10)} \quad \text{по} \quad \left\{b_{rij}^p\right\}_{r=1}^m, i\ne j, \end{cases} \tag{12}$$

через Γ_η, $\eta>0$, – подмножество номеров признаков из $X(n)$. Пошаговая реализация алгоритма иерархической агломеративной группировки будет такой:

1-й шаг: $p=0$, $\lambda c=0$, $\eta=1$. Выполнять $\Gamma_\eta=\{\eta\}$, $Margin_\eta=-2$, $\eta=\eta+1$, пока $\eta\le n$;

2-й шаг: Вычислить значения элементов матрицы $\{z_{ij}^p\}_{i,j\in I}$ по (12);

3-й шаг: Выделить $\Phi=\{z_{uv}^p \mid z_{uv}^p\ge \max(w_u^p, w_v^p)$ and $u\ne v, u,v\in I\}$. Если $\Phi=\varnothing$, то идти 9;

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

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

5-й шаг: Если $\lambda n>\lambda c$ или $\lambda n=\lambda c$ и $Margin_i<\varphi(p,i,j)$, то $\Gamma_i=\Gamma_i\cup\Gamma_j$, $\Gamma_j=\varnothing$, $Margin_i=\varphi(p,i,j)$, идти 7;

6-й шаг: Вывод номеров признаков из Γ_i, $\Gamma_i=\varnothing$, $I=I\setminus\{i\}$, идти 3;

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

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

8-й шаг: Для каждой пары (u,v), $u,v \in I$, определить значение

$$z_{uv}^{p} = \begin{cases} z_{uv}^{p-1}, u \in I \setminus \{k\}, v \in I, \\ \text{значению} \quad (10) \quad \text{на} \quad \left\{a_{rv}^{p}\right\}_{r=1}^{m}, u = k, v \in I. \end{cases}$$

Если $n–p > 1$, то идти 3;

9-й шаг: Конец.

Реализация изложенного выше алгоритма представляет один из способов решения задачи компактного описания (агрегирования) данных через поиск функциональных зависимостей между признаками. Агрегирование данных выражается в формировании нового набора из латентных признаков в описании объекта, позволяющего, не прибегая к перебору, относительно легко обнаруживать устойчивые логические закономерности.

Пусть на базе группы Γ_r, $1 \le r$, получен латентный признак $z(S)$ с максимальным значением (10), определены границы его интервалов $[c_1,c_2]$, $(c_2,c_3]$. Будем считать элементы Γ_r номерами информативного набора исходных признаков при выборе нового пространства в описании объекта $S_d \in K_t$. Значение оценки объекта $S_d \in K_t$ по латентному признаку $z(S)$ вычисляется как

$$\Omega(z(S_d)) = \theta_1(1–\theta_2), \tag{13}$$

где

$$\theta_1 = \frac{\left|\left\{S_i \in K_t \middle| z(S_i) \in [c_1,c_2]\right\}\right|}{\left|K_t\right|}, \ \theta_2 = \frac{\left|\left\{S_i \in K_{3-t} \middle| z(S_i) \in [c_1,c_2]\right\}\right|}{\left|K_{3-t}\right|}.$$

3. Вычислительный эксперимент

Для вычислительного эксперимента использовалась выборка данных по пневмококковому и серозному менингиту [5]. Каждый из 64 объектов выборки описывался 3 количественными и 18 номинальными признаками. Первый класс K_1 (пневмококковый менингит) представлен 35 объектами, второй класс K_2 (серозный менингит) – 29 объектами.

Наборы признаков для ряда объектов выборки E_0, формируемые алгоритмом пошагового включения информативных признаков по аддитивному (3) и мультипликативному (7) критериям, приводятся в табл. 1.

Таблица 1

Информативные наборы признаков объектов

№ объекта (класс)	Информативные признаки по критерию	
	аддитивному (3)	мультипликативному (7)
1(1)	$x_3, x_4, x_{12}, x_{15}, x_{16}, x_{20}$	$x_3, x_4, x_6, x_{12}, x_{15}, x_{16}, x_{17}, x_{20}$
6(1)	x_2, x_3, x_{16}, x_{20}	x_2, x_3, x_{16}, x_{20}
15(1)	x_4, x_5, x_{14}	$x_1, x_6, x_7, x_{19}, x_{21}$
37(2)	$x_3, x_7, x_{12}, x_{16}, x_{20}$	$x_2, x_3, x_7, x_{12}, x_{16}$
54(2)	x_2, x_6, x_7, x_{12}	x_2, x_6, x_7, x_{12}
57(2)	$x_2, x_3, x_4, x_6, x_7, x_{12}, x_{20}$	$x_6, x_7, x_8, x_{12}, x_{16}, x_{20}$

Как видно из табл. 1, количество признаков в наборах и их состав, определяемые по (3) и (7), сильно не отличаются друг от друга. Насколько значения оценок объектов, полученные по экстремальным значениям трёх критериев (3), (7), (8), близки друг к другу, показано в табл. 2.

Результаты экспериментов наглядно демонстрируют наличие закономерностей, выражаемых через сходство их составов информативных наборов (см. табл. 1) и близость оценок признаков объектов (см. табл. 2). Низкие показатели оценок объекта № 15 указывают на несоответствие (аномальность) значений признаков в его описании относительно представителей класса K_1.

Эффект от предобработки данных можно получить путём отказа от просмотра вариантов, не ведущих к оптимальному, с точки зрения используемого критерия, результату. Требуется проверить зави-

симость оценок объекта с учётом предобработки данных по (4). Результаты вычисления оценок по аддитивному критерию (3) на $X(n–r)$, $0 \le r \le n–2$, после удаления (справа налево) r признаков из (5) демонстрируются в табл. 3.

Таблица 2

Оценки объектов по критериям

№ объекта (класс)	Критерий		
	аддитивный (3)	мультипликативный (7)	устойчивости лог. закономерности (8)
1(1)	0,8512	0,8276	0,8000
6(1)	0,7941	0,8000	0,7143
15(1)	0,3586	0,0985	0,2571
37(2)	0,8798	0,8778	0,6897
54(2)	0,9025	0,9044	0,5172
57(2)	0,9429	0,9103	0,6897

Таблица 3

Оценки объектов по (3) с учётом предобработки

№ объекта (класс)	Число удаляемых признаков r =		
	5	10	15
1(1)	0,8571	0,8571	0,8227
6(1)	0,7941	0,7941	0,7941
15(1)	0,3586	0,3586	0,3419
37(2)	0,8798	0,8798	0,8453
54(2)	0,9025	0,9025	0,9143
57(2)	0,9429	0,9429	0,9143

Анализ результатов отбора информативных признаков и соответствующих им оценок по (3) на $X(21)$ (см. табл. 2) и с учётом предобработки на $X(16)$, $X(11)$ (см. табл. 3) показывает целесообразность поиска закономерностей в данных для сокращения комбинаторной сложности алгоритмов. Попытки применить подобный способ предобработки для вычисления информативных наборов признаков и оценок объектов по (7) и (8) оказались неэффективными.

Эффект от использования нелинейных преобразований признаков рассмотрим на примере вычисления оценок по (13). Вычисление проводилось по латентному признаку с максимальным значением (10) и соответствующему ему набору исходных признаков. Результаты представлены в табл. 4.

Таблица 4

Нелинейные отображения групп признаков на числовую ось

№ объекта (класс)	Число групп	Латентный признак получен из набора	Значение критерия (13)
1(1)	4	x_3, x_5, x_6, x_7, x_8, x_9,x_{12}, x_{15}, x_{16}, x_{20}	0,9103
6(1)	6	x_3, x_7, x_{16}	0,8246
15(1)	5	x_2, x_4, x_5, x_6, x_{10}	0,8374
37(2)	6	x_4, x_7, x_8, x_{12}, x_{17}, x_{20}	0,8571
54(2)	5	x_2, x_6, x_7, x_{10}	0,8621
57(2)	5	x_2, x_4, x_5, x_6, x_7, x_{12},x_{20}	0,9429

Неожиданный результат как эффект от использования методов интеллектуального анализа данных получен для объекта № 15. Высокая относительно критериев (3), (7), (8) оценка 0,8374 за объект по (13) свидетельствует о наличии скрытых закономерностей, обнаружить которые удаётся лишь с учётом нелинейности. Существует возможность аналитического описания закономерностей по результатам нелинейного отображения по группе признаков. Последовательность формирования латентного признака алгоритмом агломеративной иерархической группировки на примере объекта № 54 (см. табл. 4) такова:

$$x_2^1 = 0,3\left(0,0051\left(x_2^0 - 3\right) - 0,536x_6^0\right) + 0,007x_2^0x_6^0;$$

$$x_2^2 = 0,1\left(1,6813\left(x_2^1 + 0,0175\right) - 0,2426x_{10}^1\right) + 1,7133\left(x_2^1 x_{10}^1 + 0,0175\right);$$

$$x_2^3 = 1,0118\left(x_2^2 + 0,0242\right) - 0,2144x_7^2.$$

Заключение

Предложены критерии для оценки объектов по индивидуальным информативным наборам признаков. С помощью вычислительного эксперимента доказано, что, несмотря на различие по составам наборов признаков, значения оценок по различным критериям оказались близки друг к другу. Результаты вычислений могут быть использованы для наполнения баз знаний и построения информационных моделей в слабоформализованных предметных областях.

ЛИТЕРАТУРА

1. Дюк В.А. Методология поиска логических закономерностей в предметной области с нечеткой системологией: На примере клинико-экспериментальных исследований : дис. ... д-ра тех. наук. СПб., 2005. 309 с.
2. Згуральская Е.Н. Выбор информативных признаков для решения задач классификации с помощью искусственных нейронных сетей // Нейрокомпьютеры: разработка, применение. 2012. № 2. С. 20–27.
3. 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.
4. Игнатьев Н.А. Вычисление обобщённых оценок объектов и иерархическая группировка признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2015. № 4 (33). С. 31–37.
5. Касымова Р.И. Клинико-лабораторные особенности острых гнойных и серозных менингитов в зависимости от этиологии : дис. ... канд. мед. наук. Ташкент, 2009. 145 с.

Игнатьев Николай Александрович, д-р физ.-мат. наук, профессор. E-mail: ignatev@rambler.ru
Национальный университет Узбекистана. Ташкент.

Поступила в редакцию 2 июля 2016 г.

Ignat'ev Nikolay A. (National University of Uzbekistan. Republic of Uzbekistan).

Indexing of objects on the individual sets of informing features.

Keywords: patterns recognition; index of object; logical regularity; informing features of object.

DOI: 10.17223/19988605/37/3

Criteria are investigated for the estimation (indexing) of object on the individual set of different-type features. It is considered, that a training sets of $E_0=\{S_1,\ldots,S_m\}$ divided into of disjoint subset (classes) K_1,K_2. The objects of selection are described by a set of different-type features of $X(n)$, distances between objects are calculated on the metric of Juravlev. The individual set of informing features of possible object allows:

- to distinguish logical regularities in his neighborhood and calculate their stability;
- to explain a decision-making process at recognition;
- to determine belonging to the noise objects of classes;
- to produce the choice of supporting sets of features in the models of algorithms of recognition.

For the search of informing set of features of $X(k) \subset X(n)$, $k < n$ is used criteria: additive, multiplicative and to stability of logical regularity in the neighborhood of object of $S \in E_0$. The value of estimation of object of $S \in E_0$ is calculated as extremum of criterion on the individual set of $X(k)$. A heuristic allowing to decrease combinatory complication of algorithm of extraction of features on an additive criterion offers.

The method of synthesis of own space is worked out from latent features for description of object of $S \in E_0$ on the basis of hierarchical clustering. Every latent feature is result of nonlinear mapping of group of initial features on the real axis. The sequence of forming of latent feature the algorithm of hierarchical clustering is brought around to an example:

$$x_2^1 = 0,3\left(0,0051\left(x_2^0 - 3\right) - 0,536x_6^0\right) + 0,007x_2^0 x_6^0;$$

$$x_2^2 = 0,1\left(1,6813\left(x_2^1 + 0,0175\right) - 0,2426x_{10}^1\right) + 1,7133\left(x_2^1 x_{10}^1 + 0,0175\right);$$

$$x_2^3 = 1,0118\left(x_2^2 + 0,0242\right) - 0,2144x_7^2.$$

The estimations of objects can be used for the construction of models in weakly formalizing subject domains. A requirement in using on the individual set of informing features of object for making decision arises up at raising of diagnosis of illness, to development of measures on prevention of technogenic catastrophes on concrete territory. Values of estimations in [0,1] are easily interpreted in terms of fuzzy logic and can be used for verbalization of knowledge.

REFERENCES

1. Duyk, V.A. (2005) *Metodologiya poiska logicheskikh zakonomernostey v predmetnoy oblasti s nechetkoy sistemologiey: Na primere kliniko-eksperimental'nykh issledovaniy* [Methodology of search of logical regularity in a problem domain with fuzzy systemology: On the example of clinic – experimental researches]. Engineering Doc. Diss. St. Petersburg.
2. Zguralskaya, E.N. (2012) Selecting informative features for solving problems of classification using artificial neural networks. *Ney-rokomp'yutery: razrabotka, primenenie*. 2. pp. 20-27. (In Russian).
3. 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
4. Ignatiev, N.A. (2015) Computation generalized estimates of objects and hierarchical clustering of features. *Vestnik Tomskogo gosu-darstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 4(33). pp. 31-37. (In Russian).
5. Kasimova, R.I. (2009) *Kliniko-laboratornye osobennosti ostrykh gnoynykh i seroznykh meningitov v zavisimosti ot etiologii* [Clinic - laboratorial peculiarities of sharp, festering and serosal meningitises depending on etiology]. Medicine Cand. Diss. Tashkent.