ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2015 Управление, вычислительная техника и информатика № 4 (33)

УДК 519.95
DOI: 10.17223/19988605/33/4

Н.А. Игнатьев

ВЫЧИСЛЕНИЕ ОБОБЩЕННЫХ ОЦЕНОК И ИЕРАРХИЧЕСКАЯ ГРУППИРОВКА ПРИЗНАКОВ

Рассматривается процесс формирования нового признакового пространства, размерность которого меньше исходного. Предлагается последовательный отбор непересекающихся подмножеств разнотипных признаков в описании объектов и нелинейное отображения их на числовую ось. При отборе используется правило иерархической группировки для попарного объединения признаков. Решение принимается по значениям степени размытости результатов отображения объектов классов на числовой оси.
Ключевые слова: обобщённые оценки; иерархическая группировка; логические закономерности; отступ.

Обобщённые оценки – это агрегированные (комбинированные) показатели, которые в [1] использовались для отображения отношений между объектами двух классов в разнотипном признаковом пространстве на числовую ось. Было разработано два метода вычисления оценок: стохастический и детерминистический. Критерием для выбора параметров алгоритма стохастического метода служила максимальная разность (отступ) между линейными проекциями двух объектов из разных классов. Из минимального значения на числовой оси одного класса вычиталось максимальное значение другого класса. Одним из применений метода было отображение описаний (визуализация) объектов [2] на плоскость.

В алгоритме детерминистического метода применялось разбиение на интервалы доминирования значений количественных признаков объектов одного из двух классов. При вычислении обобщённой оценки объекта использовались значения функций принадлежности к интервалам доминирования для количественных признаков и частоты встречаемости градаций для номинальных признаков.

Переход к однотипным шкалам измерений и поэтапное сокращение размерности признакового пространства посредством вычисления обобщённых оценок объектов описан в [3]. На первом этапе обобщённая оценка объекта по номинальным признакам интерпретировалась как значение нового (латентного) количественного признака. На втором этапе вычисление оценки производилось по расширенному множеству количественных признаков.

Результаты вычислительного эксперимента в [3] по выборке данных GERMAN из [4] показали, что обобщающая способность решающих правил на основе обобщённых оценок выше, чем у известного метода LDA [5].

Потребность во введении латентных признаков возникает при поиске спрямляющего пространства, в котором объекты из разных классов были бы линейно разделимы. В методе опорных векторов SVM [6] нелинейность разделяющей поверхности достигается за счёт использования ядерных функций, поиск параметров дискриминантных функций производится путём максимизации отступа между объектами двух классов в новом (спрямляющем) признаковом пространстве.

В данном исследовании предлагается правило для агломеративной иерархической группировки разнотипных признаков с целью нелинейного отображения их значений в описании объектов на числовую ось. Результаты нелинейного отображения рассматриваются как значения обобщённых оценок (новых признаков) в описании объектов. Предложены критерии, на основе которых определяются число обобщённых оценок (непересекающихся групп), количество исходных признаков, входящих в группу, и их состав.

Решающие правила по значениям каждого нового признака в описании объектов образуют совокупность базовых алгоритмов. Базовый алгоритм может рассматриваться как самостоятельный классификатор либо использоваться в композиции с другими алгоритмами.

Вычисление обобщенных оценок с помощью иерархической агломеративной группировки целесообразно по нескольким причинам:

– обобщённые оценки образуют новое признаковое пространство, размеры которого меньше исходного;

– решается проблема использования алгоритмов классификации, реализация которых была неэффективна из-за большой размерности признакового пространства либо возможна при одном типе шкал измерений;

– в процессе группировки происходит последовательный отбор информативных наборов признаков;

– нелинейное отображение описаний объектов на числовую ось по определяемым комбинациям признаков является средством обнаружения устойчивых логических закономерностей (новых знаний) в хранилищах данных.

1. Обобщенные оценки объектов на базе иерархической группировки признаков

Рассматривается множество из T допустимых объектов, разбитое на 2 непересекающихся подмножества (класса). Представители классов K_1, K_2 заданы через выборку (подмножество T) объектов $E_0 = \{S_1,\ldots,S_m\}$, $E_0 = K_1 \cup K_2$. Объекты выборки описываются с помощью n разнотипных признаков $X(n) = (x_1,\ldots,x_n)$, множество допустимых значений ξ из которых измеряются в интервальных шкалах, $n - \xi$ – в номинальной.

На E_0 задано правило последовательного разбиения набора $X(n)$ на непересекающиеся подмножества $X_1(k_1),\ldots, X_\tau(k_\tau)$, $\tau \geq 1$, $k_1+\ldots+k_\tau \leq n$. Требуется для каждого $X_i(k_i)$ определить алгоритм A_i (распознающий оператор в терминологии алгебраического подхода к распознавания образов Ю.И. Журавлёва [7]) для отображения значений признаков из $X_i(k_i)$ в описании объекта $S_j \in E_0$, $j = 1,\ldots,m$, в значение (обобщённую оценку) на числовой оси.

Обозначим множество номеров количественных и номинальных признаков соответственно как I и J. Процесс последовательного вычисления значений обобщённых оценок (новых признаков) реализуется алгоритмом иерархической агломеративной группировки по описываемому ниже правилу. Для идентификации признаков в описании объектов на p-м шаге $0 \leq p < n$ иерархической группировки будем использовать $\{x_i^p\}_{i \in (I \cup J)}$.

В процессе группировки и формирования обобщённых оценок состав элементов и мощность множеств I и J, $|I| + |J| \leq n$ будут изменятся. В зависимости от шкал измерений признаков, объединяемых в группы, используются различные способы вычисления их параметров для отображения на числовую ось. Для количественных признаков это производится следующим образом.

Упорядоченное множество значений признака $x_j^p, j \in I, p \geq 0,$ объектов из E_0 разделим на два интервала $[c_1^{ip}, c_2^{ip}], (c_2^{ip}, c_3^{ip}]$, каждый из которых рассматривается как градация номинального признака. Критерий для определения границы c_2^{ip} 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^p, $j \in I,$ класса K_i, $i = 1,2$, соответственно в интервалах $[c_1^{ip}, c_2^{ip}], (c_2^{ip}, c_3^{ip}]$; $|K_i| > 1$, ν – порядковый номер элемента упорядоченной по возрастанию последовательности $r_{j_1},\ldots,r_{j_\nu},\ldots,r_{j_m}$ значений x_j^p из E_0, определяющий границы интервалов как $c_1^{jp} = r_{j_1}$, $c_2^{jp} = r_{j_\nu}$ $c_3^{jp} = r_{j_m}$. Критерий

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

позволяет оценивать значение границы между интервалами $[c_1^{ip}, c_2^{ip}]$,$(c_2^{ip}, c_3^{ip}]$. Выражение в левых скобках (1) представляет внутриклассовое сходство, в правых – межклассовое различие.

Экстремум критерия (1) 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.

При включении в группу номинального признака с целью вычисления обобщённой оценки объектов требуется определить значение его веса и вкладов каждой из градаций.

Обозначим через π число градаций признака x_r^p, $r \in J$, $p = 0$, g_{dr}^t – количество значений t-й $(1 \le t \le \pi)$ 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

$$\lambda_r = 1 - \frac{\sum_{t=1}^{\pi} g_{1r}^t g_{2r}^t}{|K_1||K_2|}. \tag{2}$$

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

$$D_{dr} = \begin{cases} \left(|K_d| - l_{dr} + 1\right)\left(|K_d| - l_{dr}\right), \pi > 2, \\ |K_d|\left(|K_d| - 1\right), \pi \le 2; \end{cases}$$

$$\beta_r = \begin{cases} \dfrac{\sum_{t=1}^{\pi} g_{1r}^t \left(g_{1r}^t - 1\right) + g_{2r}^t \left(g_{2r}^t - 1\right)}{D_{1r} + D_{2r}}, D_{1r} + D_{2r} > 0, \\ 0, \ D_{1r} + D_{2r} = 0. \end{cases} \tag{3}$$

С помощью (2), (3) вес номинального признака с $r \in J$ определяется как

$$\nu_r = \lambda_r \beta_r.$$

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

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

где $\alpha_{ij}^1, \alpha_{ij}^2$ – количество значений j-й градации i-го признака соответственно в классах K_1 и K_2, ν_i – вес i-го признака.

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

$$b_{rij}^p = \begin{cases} \mu_i\left(a_{ri}^p\right) + \mu_j\left(a_{rj}^p\right), i, j \in J, \\ \mu_i\left(a_{ri}^p\right) + t_j w_j^p \left(a_{rj}^p - c_2^{jp}\right) / \left(c_3^{jp} - c_1^{jp}\right), i \in J, j \in I, t_j \in \{-1,1\}, \\ \eta_{ij}\left(t_i w_i^p \left(a_{ri}^p - c_2^{ip}\right) / \left(c_3^{ip} - c_1^{ip}\right) + t_j w_j^p \left(a_{rj}^p - c_2^{jp}\right) / \left(c_3^{jp} - c_1^{jp}\right)\right) + \\ + \left(1 - \eta_{ij}\right) t_{ij} w_{ij}^p \left(a_{ri}^p a_{rj}^p - c_2^{ijp}\right) / \left(c_3^{ijp} - c_1^{ijp}\right), \\ i, j \in I, t_{ij}, t_i, t_j \in \{-1,1\}, \eta_{ij} \in [0,1], \end{cases} \tag{4}$$

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

$$\varphi(p,i,j)=\frac{\min\limits_{S_r\in K_1} b_{rij}^p-\max\limits_{S_r\in K_2} b_{rij}^p}{\max\limits_{S_r\in E_0} b_{rij}^p-\min\limits_{S_r\in E_0} b_{rij}^p}=\max_{t_{ij},t_i,t_j\in\{-1,1\},\eta_{ij}\in[0,1]}. \tag{5}$$

Значение (5) интерпретируется как отступ между объектами классов K_1 и K_2.

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

$$z_{ij}^p=\begin{cases}0, i=j,\\ \text{значению}\quad (1)\quad \text{по}\quad \left\{b_{rij}^p\right\}_{r=1}^m, i\ne j,\end{cases} \tag{6}$$

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

1-й шаг: $\eta=1$, $G_\eta=\varnothing$, $p=0$, $\lambda c=0$;

2-й шаг: Вычислить значения элементов матрицы $\left\{z_{ij}^p\right\}_{i,j\in(I\cup J)}$ по (6);

3-й шаг: Вычислить $\lambda n=\max\limits_{u,v\in(I\cup J)} z_{uv}^p$. Выделить $\Omega=\left\{(s,t),s,t\in I\cup J\middle| z_{st}^p=\lambda n \quad and \quad s<t\right\}$. Определить пару $\{i,j\}, i<j$, как

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

4-й шаг: Если $G_\eta=\varnothing$, то $G_\eta=\{i,j\}$, $Margin=\varphi(p,i,j)$, идти 8;

5-й шаг: Если $G_\eta\cap\{i,j\}=\varnothing$, то идти 7;

6-й шаг: Если $\lambda n>\lambda c$ или $\lambda n>\lambda c$ и $Margin<\varphi(p,i,j)$, то $G_\eta=G_\eta\cup\{i,j\}$, $Margin=\varphi(p,i,j)$, идти 8;

7-й шаг: $\eta=\eta+1$, $G_\eta=\varnothing$. Идти 4;

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

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

9-й шаг: Определить значение

$$z_{uv}^p=\begin{cases}z_{uv}^{p1}, u\in I\setminus\{k\}, v\in I,\\ \text{значению}\quad (1)\quad \text{по}\quad \left\{a_{rv}^p\right\}_{r=1}^m, u=k, v\in I.\end{cases}$$

Если $n-p>1$, то идти 3;

10-й шаг: Конец.

Через конечное число рекурсивных обращений к описанному выше алгоритму все исходные признаки сводятся к одной нелинейной оценке. По практическим соображениям ограничение на число обобщённых оценок для конкретных выборок данных может определяться по результатам вычислительного эксперимента либо исходя из дополнительных критериев выбора.

Рассмотрим пример классификатора на базе обобщённых оценок (4). Пусть $\left\{a_{ir}^p\right\}_{i=1}^m$, $p<n$, $r\in I$ – множество значений обобщённой оценки (признака), вычисленной по (4), и по критерию (1) эти значения разбиты на интервалы $[c_1,c_2],(c_2,c_3]$. Для решающего правила нужно выбрать порог, равный

$$w_0=\frac{c_2+z}{2}, \tag{7}$$

где z – ближайшее к c_2 значение из интервала $(c_2,c_3]$. Анализ результатов использования порога (7) в дискриминантных функциях приводится в [3].

3. Вычислительный эксперимент

В качестве материала для эксперимента была взята выборка данных из [8], описывающая челюсти 30 собак (класс K_1) и 12 волков (класс K_2) по следующим 6 количественным признакам:

x_1 – (CBL) основная длина;

x_2 – (LUJ) длина верхней челюсти;

x_3 – (WID) ширина верхней челюсти;

x_4 – (LUC) длина верхнего карнивора;

x_5 – (LFM) длина первого верхнего моляра;

x_6 – (WFM) ширина первого верхнего моляра.

Порядок синтеза значений обобщённых оценок (латентных признаков) из комбинаций признаков по критерию (1) и отступов между объектами классов по (5) приведён в табл. 1.

Таблица 1

Порядок синтеза обобщённых оценок объектов

Комбинация признаков	Значение критерия (1)	Отступ между классами (5)
X_1, x_4	1,0000	0,0403
x_1, x_4, x_5	1,0000	0,1060
x_1, x_4, x_5, x_3	1,0000	0,1233
x_1, x_4, x_5, x_3, x_2	1,0000	0,1674
x_1, x_4, x_5, x_3, x_2, x_6	1,0000	0,1778

Аналитический вид решающего правила по значениям обобщённой оценки, полученной при синтезе признаков x_1 и x_4 (табл. 1) с учётом (7), будет выглядеть так:

$$d(x) = 0{,}4(-0{,}0037(x_1-221)-0{,}09538(x_4-22{,}5))+0{,}0001(x_1x_4-5130)+0{,}01971.$$

Судя по результатам табл. 1, все комбинации исходных признаков попадают в одну группу, попарное объединение признаков в комбинацию удовлетворяет такому свойству, как монотонность по значениям отступа между объектами классов. Теоретическое обоснование выполнения монотонности при синтезе комбинаций признаков на произвольной двухклассовой обучающей выборке требует отдельного рассмотрения. Возможным вариантом решения проблемы монотонности является обнаружение и исключение из выборки шумовых объектов.

Для демонстрации того, что различные признаки в составе обобщённых оценок (4) компенсируют недостатки друг друга, воспользуемся табл. 2 [1]. Таблица содержит значения границ интервалов $[c_1^i, c_2^i], (c_2^i, c_3^i]$ и экстремумы критерия (1) для признаков $\{x_i\}$, $i = 1,\ldots,6$.

Таблица 2

Результаты оптимизации по критерию (1)

Признак	c_1^i	c_2^i	c_3^i	w_i
x_1	129,000	221,000	255,000	0,378
x_2	64,000	114,000	126,000	0,389
x_3	52,000	76,000	95,000	0,288
x_4	16,700	22,500	26,500	0,897
x_5	11,200	14,700	16,800	0,625
x_6	13,000	18,300	27,000	0,800

Согласно [1] точность классификации по линейным дискриминантным функциям (дискриминанта Фишера в том числе) напрямую зависит от использования признаков x_4 и x_6, имеющих наибольшие значения весов (табл. 2), равных соответственно 0,897 и 0,800. Доказано, что корректное разделение линейных проекций объектов обучения на классы с единичным значением критерия (1) возможно лишь на наборах $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ и $\{x_1, x_2, x_4, x_5, x_6\}$. Наилучший результат (см. табл. 1) в смысле разделимости по (1) и отступа между классами (5) по различным парам исходных признаков был достигнут при использовании нелинейного отображения значений из $\{x_1, x_4\}$ в описании объектов на числовую ось.

С помощью критерия (1) преобразуем количественные признаки в номинальные. Каждой градации номинального признака поставим в соответствие один из непересекающихся интервалов, полученный по (1). В табл. 3 приведены результаты группировки для случая, когда все признаки номинальные,

в табл. 4 представлено два подмножества: $\{x_1, x_2\}$ – количественных и $\{x_3, x_4, x_5, x_6\}$ – номинальных признаков.

Таблица 3

Группировка по номинальным признакам

№ группы	Состав группы	Значение критерия (1)
1	x_2, x_4, x_5, x_6	0,8965
2	x_1x_3	0,3781

Таблица 4

Группировка по разнотипным признакам

№ группы	Состав группы	Значение критерия (1)
1	x_1, x_2, x_4, x_5	1,0000
2	x_3,$x6$	0,2884

Анализ результатов из табл. 1, 2 и 3 по критерию (1) показывает, что преобразование значений признаков из количественных (сильных) шкал измерений в значения номинальной (слабой) шкалы приводят к снижению точности решающих правил с порогом (7) на базе обобщённых оценок.

Для проверки процедурой кросс-валидации обобщающей способности решающих правил с порогом (7) на базе нелинейных обобщённых оценок с максимальным отступом между классами использовалось разделение выборки на обучение и контроль в соотношении 9:1. Результаты проверки следующие: точность на обучении – 100%, на контроле – 98%.

Заключение

Процесс вычисления обобщённых оценок сводится к формированию нового признакового пространства для описания допустимых объектов в задачах распознавания образов. Практическое применение этих оценок позволяет:

– находить устойчивые логические закономерности в базах (хранилищах) данных, не прибегая к перебору всевозможных вариантов;

– использовать их для реализации дискриминантных функций, решающих списков, решающих деревьев, алгоритмов вычисления оценок.

Теоретический и практический интерес представляет оценка границ допустимых значений латентных признаков на основе обобщённых оценок.

ЛИТЕРАТУРА

1. *Игнатьев Н.А.* Вычисление обобщённых показателей и интеллектуальный анализ данных // Автоматика и телемеханика. 2011. № 5. С. 183–190.
2. *Игнатьев Н.А.* О конструировании признакового пространства для поиска логических закономерностей в задачах распознавания образов // Вычислительные технологии. 2012. Т. 17, № 4. С. 56–62.
3. *Игнатьев Н.А., Нуржонов Ш.Ю.* Выбор параметров регуляризации для повышения обобщающей способности дискриминантных функций // Узбекистон Республикаси Курол Кучлари академиясининг хабарлари. 2014. № 1 (14). С. 81–87.
4. *Asuncion A., Newman D.J.* UCI Machine Learning Repository // University of California, Irvine. 2007. URL: www.ics.uci.edu/mlearn/MLRepository.html
5. URL: http://www.mathworks.com/help/stats/discriminant-analysis.html
6. *Потапов А.С.* Технологии искусственного интеллекта. СПб. : СПбГУ ИТМО, 2010. 218 с.
7. *Журавлёв Ю.И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. М. : Наука, 1978. Т. 33. С. 5–68.
8. *Жамбю М.* Иерархический кластер-анализ и соответствия. М. : Финансы и статистика, 1988. 342 с.

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

Поступила в редакцию 22 сентября 2015 г.

Ignat'ev Nikolai A. (National University of Uzbekistan. Republic of Uzbekistan).
Computation generalized estimates of objects and hierarchical clustering of features
Keywords: generalized estimates; hierarchical clustering; logical regularity; margin.

DOI: 10.17223/19988605/33/4

We consider the set acceptable objects in T, broken into two disjoint subsets (classes). Representatives of the classes K_1, K_2 are given by the sample objects $E_0 = \{S_1,\ldots,S_m\}$, $E_0 = K_1 \cup K_2$. Objects of the sample are described by n heterogeneous features $X(n) = (x_1,\ldots,x_n)$ the set of acceptable values of ξ which measurements are on the interval scale and n - ξ are on the nominal.

Given the rule is a sequence of partitions set $X(n)$ into disjoint subsets $X_1(k_1),\ldots,X_\tau(k_\tau)$, $\tau \geq 1$, $k_1+\ldots+k_\tau \leq n$ at E_0. Required for each $X_i(k_i)$ algorithm to determine A_i nonlinear display feature values $X_i(k_i)$ of the object in the description $S_j \in E_0$, $j=1,\ldots,m$, a value (generalized estimation) on the real axis.

For the identification generalized estimates (new features) in object descriptions on p-th step $0 \leq p < n$ hierarchical clustering is used $\{x_i^p\}_{i \in (I \cup J)}$, where I and J, respectively, the set of indices of nominal and quantitative features. Making a decision to merge pairs of features is based on the interval analysis of the results of non-linear displaying them on the real axis and the margin between classes.

The ordered set feature values of objects in E0 is divided into two intervals $[c_1^{jp}, c_2^{jp}],(c_2^{jp}, c_3^{jp}]$. The criterion for determining the borders c_2^{jp} based on hypothesis testing (assertion) that each of the two intervals contains the values of a quantitative feature of only one class objects.

Let u_i^1, u_i^2 – the number of characteristic values $x_j^p, j \in I$ class K_i, $i = 1,2$ respectively, in the intervals $[c_1^{jp}, c_2^{jp}],(c_2^{jp}, c_3^{jp}]$, $|K_i| > 1$, ν – the ordinal number of the element ordered ascending sequence of x_j^p values of E_0, determining interval limits as $r_{j_1},\ldots,r_{j_\nu},\ldots,r_{j_m}$. Criterion

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

allows to assess the meaning borders between interval $[c_1^{jp}, c_2^{jp}],(c_2^{jp}, c_3^{jp}]$. The extremum of the criterion (1) is used as a weight w_j^p ($0 \leq w_j^p \leq 1$) of feature x_j^p and for a decision by the rule of hierarchical clustering.

If $\forall x_t \in X(n) \setminus \bigcup_{d=1}^{i} X_d(k_d)$ $\left|\bigcup_{d=1}^{i} X_d(k_d)\right| < n$ nonlinear mapping on the real axis of objects E_0 on $X_i(k_i) \cup \{x_t\}$ value (1) less than or equal to the analogical value (with a less margin between classes) on $X_i(k_i) \cup \{x_t\}$ that is formed new group for the synthesis of the generalized estimation.

Calculation of generalized estimates using hierarchical clustering advisable for several reasons:

– generalized estimates form a new feature space whose dimensions are smaller than the original;

– solves the problem of the use of classification algorithms, the implementation of which was inefficient due to the large dimension of feature space, is possible at any single type measurement scales;

– in the process of clustering occurs consistent selection of informative feature sets;

– nonlinear mapping object description to real axis defined by a combination of features is a means of detection stable patterns of logic (new knowledge) in data warehouses.

REFERENCES

1. Ignat'ev, N.A. (2011) Computing generalized parameters and data mining. *Automation and Remote Control*. 72 (5). pp. 183-190. DOI: 10.1134/S0005117911050146
2. Ignat'ev, N.A. (2012) On the construction of the feature space for finding logical regularities in pattern recognition problems. *Vychislitel'nye tekhnologii – Computational Technologies*. 17(4). pp. 56-62. (In Russian).
3. Ignat'ev, N.A. & Nurzhonov, Sh.Yu. (2014) Vybor parametrov regulyarizatsii dlya povysheniya obobshchayushchey sposobnosti diskriminantnykh funktsiy [The choice of the regularization parameters for improving the generalization ability of the discriminant functions]. *Uzbekiston Respublikasi Kurol Kuchlari akademiyasining khabarlari – Bulletin of the Academy of the Armed Forces of the Republic of Uzbekistan*. 1(14). pp. 81-87.
4. Asuncion, A. & Newman, D.J. (2007) *UCI Machine Learning Repository*. University of California, Irvine. [Online] Available from: www.ics.uci.edu/mlearn/MLRepository.html
5. Mathworks. (n.d.) *Discriminant Analysis*. [Online] Available from: http://www.mathworks.com/help/stats/discriminant-analysis.html
6. Potapov, A.S. (2010) *Tekhnologii iskusstvennogo intellekta* [Artificial intelligence technology]. St. Petersburg: SPbGU ITMO.
7. Zhuravlev, Yu.I. (1978) Ob algebraicheskom podkhode k resheniyu zadach raspoznavaniya ili klassifikatsii [On an algebraic approach to solving the problems of pattern recognition and classification]. In: Girevich., I.B. (ed.) *Problemy kibernetiki* [Problems of Cybernetics]. Vol. 33. Moscow: Nauka. pp. 5-68.
8. Jambu, M. (1988) *Ierarkhicheskiy klaster-analiz i sootvetstviya* [Hierarchical cluster analysis and compliance]. Moscow: Finansy i statistika.