IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 10, October 2016
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

Analytical Representation of Recognition Operators to Calculate the Generalized Estimation

Saidov Doniyor Yusupovich

Department of algorithms and programming technologies, National university of Uzbekistan named after Mirzo Ulugbek, Tashkent, 100123, Uzbekistan

Abstract

In this paper considered the analytical representation of recognition operators to calculate the estimations in problems of recognition. Describes a method for the formation the set of features by the rule of hierarchical agglomerative grouping and non-linear displaying them to the numerical axis. Displaying of such set to a numerical axis forms values of a latent sign in the description of admissible objects. Experimentally proved that the latent quantitative features are ordered by the degree of informativeness at hierarchical grouping.

Key words: *recognizing operator, hierarchical agglomerative grouping, non-linear displaying, analytical representation*

1.Introduction

The concepts of recognizing operator and the decision rule has been entered in the framework of algebraic approach to pattern recognition [1] for the presentation of recognition algorithm. The recognizing operator translates the descriptions of valid objects to the estimations which the decision rule classifies objects to one or another class. One of the forms to obtain estimates may be their calculation by an analytical representation of operators. Analytic representations commonly used in the implementation of the mathematical models in computer systems.

A lot of program complexes are oriented to obtain the analytical form of the functions in the differentiation and integration. In data mining a knowledge representation based on "if - then" rules also have analytical form. Conditions in "if - then" rules could be given in one of the forms of logical regularities. In most cases logical expressions are given in the form of parallelograms.

Generalized estimation - is the aggregated (combined) indexes, which are used to display the relationship between objects of the two classes in different-type feature space to the numerical axis [2]. In this paper considers nonlinear displaying the set of defined features to the numerical axis. To find these sets uses a special rule for hierarchical agglomeration grouping. The differences between linear and nonlinear methods of calculating the generalized estimations is given in Table 1.

Table 1. The differences between linear and nonlinear methods

The method of calculating the generalized estimations	
Linear	Nonlinear
Sets of features specified by the user.	Sequential selection of feature sets is produced by the rule of agglomerative hierarchical grouping.
Method of selection the informative features does not exist.	Exists the method of selection the informative features.
There is no method ordering feature sets by their compactness.	It increases the compactness of objects, classes, and the sample as a whole on defined sets of features.

To obtain an analytical representation of nonlinear displaying the defined set of features on the numerical axis is invited to apply the apparatus of formal grammar. When outputting chains of terminal symbols used downstream analysis. The process obtaining the analytic representation (formulas for calculating estimation) demonstrated in a test example.

2. Statement of the problem and the non-linear method of calculating the generalized estimation.

Two class problem of recognition in the standard formulation is considered. Each object of the sample $E_0 = \{S_1,...,S_m\}$ belongs to one of the classes K_1 or K_2 ($E_0 = K_1 \cup K_2$) and is describes by n quantitative characters $X(n) = (x_1,...,x_n)$.

There is given a rule on E_0 which sequentially divide the set $X(n)$ into disjoint subsets $X_1(k_1),...,X_\tau(k_\tau), \tau \ge 1,\ k_1 + ... k_\tau \le n$ .Requires:

– for each $X_i(k_i)$ determine an algorithm A_i (recognizing operator in terms of the algebraic approach to pattern recognition Zhuravlev Y.I. [1]) to display

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 10, October 2016
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

values of the features $X_i(\mathrm{k}_i)$ on describing object $S_j \in E_0, j=\overline{1,m}$ into value (generalized estimation) on numerical axis;

– get the analytical form (formulas) to calculate the recognizing operator A_i.

To identify the features as an initial so obtained by calculating the generalized estimation at p – th step $0 \le p < n$ of hierarchical agglomerative grouping will use the $\{x_i^p\}_{i=1}^{n-p}$. The set of quantitative features numbers will be denoted by I. If $p=0$ the number of groups coincide with the number of features in $X(\mathrm{n})$ and $I=\{1,...,\mathrm{n}\}$.

In the process of grouping and formation generalized estimation the composition of elements and cardinality of the set I, $|I| \le n$ will be change.

The ordered set of values of the feature $x_j^p, j \in I, p \ge 0$ objects from E_0 will divided into two intervals $\left[c_1^{jp}, c_2^{jp}\right], \left(c_2^{jp}, c_3^{jp}\right]$, each of which is considered as a gradation of the nominal feature.

The criterion to determining the boundaries of c_2^{jp} is based to check the hypothesis (statements) that each of the two intervals contains the quantitative feature values of objects of only one class.

Let u_i^1, u_i^2 be the number of the values of feature $x_j^p, j \in I$ from the class $K_i, i=1,2$ in the intervals $\left[c_1^{jp}, c_2^{jp}\right], \left(c_2^{jp}, c_3^{jp}\right]$, $|K_i| \ge 1$, and υ – is the sequence number of an element in the ascending ordered sequence $r_{j1},...,r_{j\upsilon},...,r_{jm}$ of the values x_j^p from E_0, which defines the boundary of the interval $c_1^{jp}=r_{j1}, c_2^{jp}=r_{j\upsilon}, c_3^{jp}=r_{jm}$. The criterion

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

makes it possible to calculate the optimal value of the boundary of the interval $\left[c_1^{jp}, c_2^{jp}\right], \left(c_2^{jp}, c_3^{jp}\right]$. The expression on the left-hand brackets (1) is the intraclass similarity, in the right - interclass difference.

Extremum of criterion (1) is used as a weight $w_j^p\left(0 \le w_j^p \le 1\right)$ of the feature x_j^p. When $w_j^p=1$, values of the feature x_j^p in objects from K_1, K_2 classes not intersect.

The value of generalized estimation b_{rij}^p of the object $S_r=\left(a_{r1}^p,...,a_{r,n-p}^p\right)$, $S_r \in E_0$ by pair of $\left(x_i^p, x_j^p\right), 1 \le p < n,\ i,j \in \{1,...,\mathrm{n}-\mathrm{p}\},\ \mathrm{i} \ne \mathrm{j}$ is calculated as

$$\begin{aligned} b_{rij}^p &= \mu_{ij}\left(t_i w_i^p\left(\frac{a_{ri}^p-c_2^{ip}}{c_3^{ip}-c_1^{ip}}\right)+t_j w_j^p\left(\frac{a_{rj}^p-c_{21}^{jp}}{c_3^{jp}-c_1^{jp}}\right)\right)+ \\ &+\left(1-\mu_{ij}\right)t_{ij} w_{ij}^p\left(\frac{a_{ri}^p a_{rj}^p-c_2^{ijp}}{c_3^{ijp}-c_1^{ijp}}\right) \\ & i,j \in I, t_i, t_j, t_{ij} \in \{-1,1\}, \mu_{ij} \in [0,1], \end{aligned} \qquad (2)$$

where w_i^p, w_j^p, w_{ij}^p - features weight defined by (1) respectively over the set of features values x_i^p, x_j^p and their multiplication $x_i^p x_j^p$, the value $t_i, t_j, t_{ij} \in \{-1,1\}, \mu_{ij} \in [0,1]$ is selected by extremum of the functional

$$\varphi(\mathrm{p,i,j}) = \frac{\min_{S_r \in K_1} b_{rij}^p - \max_{S_r \in K_2} b_{rij}^p}{\max_{S_r \in E_0} b_{rij}^p - \min_{S_r \in E_0} b_{rij}^p} \to \max. \qquad (3)$$

Value (3) is interpreted as a margin (offset) between objects of the classes K_1 and K_2. Let we denote as $\{z_{ij}^p\}_{i,j=1}^{n-p}, p \ge 0$ the matrix, which elements values z_{ij}^p is defined as

$$z_{ij}^p = \begin{cases} 0, & i=j \\ value\ of\ (1)\ on \{b_{rij}^p\}_{r=1}^m, & i \ne j, \end{cases} \qquad (4)$$

through $G_\mu, \mu > 0$ denotes a subset of the features numbers from $X(n)$. Step by step implementation algorithm of the iterative grouping will be:

Step 1: $\mu=1, G_\mu=\varnothing, p=0, \lambda c=0$;

Step 2: Calculate the values of elements of the matrix $\{z_{ij}^p\}_{i,j\in I}$ by (4)

Step 3: Calculate $\lambda n = \max_{u,v \in I} z_{uv}^p$. Choose $\Omega=\{(\mathrm{s,t}),\ \mathrm{s,t} \in \mathrm{I} \mid z_{st}^p=\lambda n\ and\ s<t\}$. Define a pair of $\{i,j\}, i<j$ as

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

Step 4: If $G_\mu=\varnothing$ then $G_\mu=\{i,j\}$, $Margin=\varphi(p,i,j)$ and go to Step 8;

Step 5: If $G_\mu \cap (i,j) = \varnothing$ then go to Step 7;

Step 6: If $\lambda n > \lambda c$ or $\lambda n = \lambda c$ and $Margin < \varphi(p,i,j)$ then $G_\mu = G_\mu \cup \{i,j\}$, $Margin=\varphi(p,i,j)$ and go to Step 8;

Step 7: $\mu=\mu+1, G_\mu=\varnothing$. Go to Step 4;

Step 8: $p = p+1$, $I = I \setminus max(i,j)$, $I = I \cup \min(i,j)$, $k = \min(i,j)$, $\lambda c = \lambda n$. Replace the values of features in description of the object $S_r = \{a_{ru}^{p-1}\}_{u \in I}$, $r = 1,...,m$ to

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

3.Analytical representation of recognition operators

To obtain an analytical representation of recognition operators used methods of the theory of formal grammars. The set of terminal symbols contain the name of initial features and parameters $\left(w_i^p, w_j^p, w_{ij}^p, t_i, t_j, t_{ij}, \mu_{ij}\right)$. The set of non-terminal symbol is represented by two elements {<formula>, <latent feature>}. Initial symbol of the grammar in this case is <formula>.

Computational experiment was conducted on medical data [3] with the parametres of hypertension. The sample of 147 objects has been divided into two classes: K_1 (healthies) contained indexes of 111 objects, K_2 (patients) - 36 objects. Each object is described by 29 features. On table. 2 presents the results of sequential association initial features in the group taking into account the nesting parentheses

Table 2. Results of grouping the features.

Group number	Sequences of association the features	The value of the criterion (1)	Margin by (3)
1	(((((((4, 20), 9), 18), 8), 2), 10),12)	1	0.0105
2	((((((5,14), 16), 19), 23), 3), 7)	1	0.0086
3	((((((((((26, 28), 27), 15), 1), 25), 13), 21), 22), 24), 11), 17)	0.9672	-0.0042
4	(6, 29)	0.7331	-0.238

The analytical form of the sequence calculating the values of the first latent feature (see table 2) with the initial features names in the "[,]" was this:

x_4^1 =-0,0069476([АДС]-140)+0,3949233([ДИАСТОЛА]-0,42)-0,0041346([АДС]*[ДИАСТОЛА]-68,2);

x_4^2 =0,8013498(x_4^1 -0,0094)+3,1287087([QRS]-0,08);

x_4^3 =0,5811987(x_4^2)+0,551857([СИСПОК]-0,485)+0,8920181(x_4^2 *[СИСПОК]);

x_4^4 =0,2117403(x_4^3 +0,0388)+0,2756011([QT]-0,36)+2,136341(x_4^3 *[QT]+0,0124);

x_2^5 =0,0057899([РОСТ]* x_4^4 +1,9623);

x_2^6 =0,3107939(x_2^5 *[ПЛП]);

x_2^7 =0,1804337(x_2^6 *[КДР]).

The value of the criterion (1), equals to 1 by the feature x_2^7 demonstrates that the nonlinear displaying the features set on the numerical axis lead to correct (without error) separating the objects of a sample into classes using only 8 (out of 29) of the original features.

4.Data analyses with description in different type feature space

The problem of selecting the analytical description of decision rules in different type feature space can be solved by converting the nominal features in quantitative. There are two alternatives to solve the problem:

– extension of the space by describing each gradation in the nominal scale as a separate feature;

– displaying defined groups nominal features on the numerical axis.

The first alternative is not acceptable as a part of the dimension of the space so and explanations (interpretations) of the obtained results. Solution by the second alternative is possible by using the technologies of calculating generalized estimates[2], given below.

Let us denote by p number of gradation of the feature $r \in J$, g_{dr}^t - number of values of the t -th ($1 \le t \le p$) gradation of r -th feature in the description objects of class K_d, l_{dr} - number of gradation of r -th feature in K_d. The difference over r -th feature between the classes K_1 and K_2 is determined as a value

$$\lambda_r = 1 - \frac{\sum_{t=1}^{p} g_{1r}^t g_{2r}^t}{|K_1||K_2|}. \quad (5)$$

The degree of uniformity (intraclass similarity measure) β_r gradation values of r-th feature by classes of K_1, K_2 is calculated according to the formulas:

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

$$\beta_r = \begin{cases} \dfrac{\sum_{t=1}^{p} g_{1r}^t (g_{1r}^t - 1) + g_{2r}^t (g_{2r}^t - 1)}{D_{1r} + D_{2r}}, D_{1r} + D_{2r} > 0, \\ 0, \qquad D_{1r} + D_{2r} = 0. \end{cases} \quad (6)$$

With (5), (6) weight of nominal feature x_r is determined as

$$v_r = \lambda_r \beta_r. \quad (7)$$

It is easy to verify that the set of weights values of nominal features, calculated by (7) belong to the interval $[0,1]$.

It is obvious that the set of numbers identifying as p gradation of the nominal feature, always possible one-to- one display into the set $\{1,...,p\}$. Taking into account such a mapping for the object $S=(x_1,...,x_n)$ contribution $x_i=j$, $i\in J$, $j\in\{1,...,p\}$ feature in generalized estimation is determined by the

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

where $\alpha_{ij}^1,\alpha_{ij}^2$ -are the numbers of the values j-th gradation of i-th feature accordingly in classes K_1 and K_2, v_i- is a weight of i-th feature, calculating by (7). In different – type features space generalized estimation for each object $S_a\in E_0$, $S_a=(x_{a1},...,x_{an})$ will be calculated as

$$R(S_a)=\sum_{i\in J}\mu_i(x_{ai}) \qquad (8)$$

Selecting latent quantitative features can be made according to the rules of hierarchical agglomerative grouping [4]. Representatives of each group are displayed on the numerical axis by (8) as a separate feature in description of objects. After such conversion to the representation of objects may be used a nonlinear algorithm described above. Unfortunately mathematical basis of selection rules for hierarchical grouping has not been developed.

To demonstrate how to work with different type data will use displaying a set of all nominal features in one latent feature on the example Statlog (heart).

The sample consists of 270 objects, of which 150 belongs to the K_1 class (no cardiovascular disease), 120 to K_2 (the presence of cardiovascular disease).

Each object is described by the set of $X(13)=(x_1,...,x_{13})$ containing 7 nominal and 6 quantitative features.

From 7 nominal features by (8) formed a latent quantitative feature under the number 7. The value of the criterion (1) and the results of grouping features taking into account the indicated transformation are shown in the tables 3,4.

Table 3. The names and weights of the features from [5] by (1).

№	Feature number	Feature name	Conditional features abbreviations	The value of the criterion (1)
1	7	Latent (combination of all nominal features)	LAT	0.5035
2	6	number of major vessels (0-3) colored by fluoroscopy	NOMV	0.3772
3	4	maximum heart rate achieved	MHRA	0.3413
4	5	old peak = ST depression induced by exercise relative to rest	OL	0.3312
5	1	age	AGE	0.2871
6	3	serum cholesterol in mg/dl	SCH	0.2685
7	2	resting blood pressure	RBP	0.2548

Table 4. The results of grouping of the features.

№	The sequence of association of features	The value of the criterion (1)	Margin by (3)
1	(((6, 7), 2), 4)	0.6064	-0.4335
2	(1,5)	0.3481	-0.6446
3	3	0.2685	-0.7682

The analytical form of convertation (see table 4) was as follows:

x_6^1 =-0.037723[NOMV]+0.105103([LAT]+0.2106)+ 0.085928([NOMV]*[LAT]+0.0117);

x_2^2 =0.008431([x_6^1]*[RBP]+0.0706);

x_2^3 =0.006607[x_2^2]*[MHRA].

References

[1] Zhuravlev Y.I. Algebraic methods in recognition problems and classification // Recognition, classification, prediction. Mathematical methods and their application, Science, 1989.- Vol. 1.- p 9-16.

[2] N. Ignatiev Generalized estimation and local metrics of the objects in data mining, Tashkent University, 2014. - 73 p.

[3] 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.

[4] Duke V.A. Methodology of finding the logical patterns in the domain of fuzzy systemology: On the example of clinical and experimental studies. St. Petersburg, 2005. - 309 p.

[5] https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)