STABILITY OF THE OBJECTS OF CLASSES AND GROUPING THE FEATURES

A.X.Xurramov
Karshi State University. Karshi , Uzbekistan
Xurramov2009@yandex.ru
https://doi.org/10.5281/zenodo.10074384

O‘qituvchi bilan anglash masalalarida yangi alomatlar fazosini tanlash masalasi qaraladi. Fazoni shakllantirish uchun boshlang‘ich alomatlar to‘plami, soni oldindan ma’lum bo‘lmagan o‘zaro kesishmaydigan guruhlarga bo‘linadi. Guruhga kiruvchi alomatlar to‘plam ostisi bo‘yicha ob’ektlarning o‘z sinfidagi turg‘unlik qiymatlari orqali guruhlash kriteriyasining ekstremumi hisoblanadi. Bunday to‘plam ostisini sonlar o‘qiga akslantirilishi, mumkin bo‘lgan ob’ektlarni tavsiflovchi latent alomat qiymatlarini ifodalaydi. Taklif qilingan guruhlash usuli informativlik bo‘yicha tartiblangan alomatlarning monoton ketma-ketligini hosil qilishi isbotlangan. Latent alomatlardan tashkil topgan yangi alomatlar fazosiga o‘tish berilganlar bazasidan yashiringan qonuniyatlarni izlash imkoniyatlarini ancha kengaytiradi.
Tayanch iboralar: ob’ektlar turg‘unligi, latent alomatlar, klasterlash, umumlashgan baholar, yashirin qonuniyatlar.

Рассматривается задача выбора нового признакового пространства в задачах распознования с учителем. Для формирования пространства используется разбиение множества исходных признаков на заранее неизвестное количество непересекающих групп. Экстремум критерия группировки вычисляется по значениям устойчивости каждого объекта в своём классе по набору признаков входящих в группу. Отображение такого набора на числовую ось представляет значения латентного признака в описании допустимых объектов. Доказано, что предложенный метод группировки формирует монотонную последовательность латентных признаков, упорядоченную по отношению информативности. Переход в новое пространство из латентных признаков позволяет значительно расширить возможности по поиску скрытых закономерностей из баз данных.
Ключевые слова: устойчивость объектов, латентные признаки, кластеризация, обобщенные оценки, скрытые закономерности.

It is considering a problem of selection a new feature space in recognition problems with teacher. For formation the space uses division the set of initial features into previously unknown number of disjoint groups. Extremum of the criterion of grouping is calculated by the stability values of each object in its class on the set of features in the group. Displaying such a set to the numerical axis represents the values of the latent feature in the description of admissible objects. it is proved that the offered method of grouping forms a monotone sequence of latent features ordered in the relation of informativeness. Transition into the new space of the latent features allows significantly expand opportunities to find a hidden regularities from the databases.
Keywords: stability of objects, latent features, clustering, generalized estimation, hidden regularities.

1. Introduction

of objects is one of the important concepts in the field of image perception. It is the main indicator to confirm the truth of the compactness hypothesis [1]. In turn, based on the compactness hypothesis, it is possible to create methods for constructing new symptom spaces.

It is known that compactness also changes as the symptom space for one sample changes. For this reason, the problems of dividing the characteristics of the selected objects into groups and researching the compactness of the objects in these groups are relevant. The stability of objects in the group of symptoms identified here shows the level of truth of the emerging laws.

Grouping of symptoms in cases where the symptom space is very large, it is possible to use standard algorithms by moving to a small-sized symptom space where new latent symptoms are formed. In this case, it will be possible to solve the problem set even within the boundaries of one group of symptoms, and finally, it will be possible to determine the relations between the groups and the internal laws hidden in the properties of the groups.

One of the methods of grouping signs around an object is the local geometry method [2]. Its difference from traditional methods is that each object of the sample is considered as an independent classifier and its own (local) symptom space is built for it. Individual dimensions of similarities and differences with other objects are determined, and on this basis logical regularities are sought in the structure of givens.

It is known that the grouping of symptoms is based on the identification of informative symptoms. In particular, the article [3] lists five types of extraction of informative symptoms of different categories: a discrete method of searching for informative zones in the training sample; clustering; based on the assumption of normal distribution of objects in clusters; based on the information-theoretic concept of entropy; methods based on nonparametric estimation of density. These methods operate based on initial parameters and probabilistic assumptions.

The article proposes a set of methods that are applicable in cases where, unlike the types listed above, the number of groups of symptoms and its shape are not known in advance. Grouping of the selected objects of symptoms is carried out on the basis of pairwise proximity relations and compactness hypothesis in the classification of objects.

In cases where the symptom space is large, it is appropriate to use a hierarchical agglomerative algorithm for grouping independent informative symptoms, as the proximity relations between objects are "washed away" under the influence of "bad" symptoms.

From a theoretical point of view, the classification of symptoms into groups is an NP problem, with the number of variants of the groups $C_n^2 + C_n^3 + \ldots + C_n^{n-1}$ equal to . In the work under consideration, the only option of groups of symptoms is found based on the criterion set in the hierarchical agglomerative grouping algorithm of symptom grouping. The number of groups and their composition are determined by grouping rules.

At work it is offered:

– selection of symptoms using the hierarchical agglomerative grouping algorithm according to the object stability criterion;

– extracting a subset of informative latent symptoms.

2. Setting the issue

The problem of understanding the standard images is considered. A set of objects containing representatives of $E_0 = \{S_1,\ldots,S_m\}$ non-intersecting classes $K_1,\ldots,K_d$ $(d>1)$ is considered given. The possible object of selection is characterized by ξn different categories of symptoms (quantitative and qualitative), $X(n)=(x_1,\ldots,x_n)$ one of which is measured in intervals and $n-\xi$ one in nominal scales. It 's medicine and we denote the sets of nominal symbol numbers t by and J respectively I , $|I|+|J|=n.$

E_0 consider $S \in E_0 \cap K_p$ the feature defined in the sample K_p to be defined functionally, which calculates the class stationarity under the set of features $U(S,X(k))$ defined in the object . $X(k),\ X(k)\subset X(n)$

, it can be considered that the sample objects are divided into two K_1 and classes K_2 without violating the property of generalization . E_0

Based on the classification of symptoms into groups, the formation of a new symptom space is required to be carried out in the form of the following steps in sequence:

– E_0 the set of symptoms into non $U(S_i,X(k))$- $l \le n$ intersecting groups using their values $G_1,\ldots,G_l$ according to the description of the objects $X(n)$ of the sample $S_i, i=1,\ldots,m$;

– $G_j,\ j=1,\ldots,l$ of the sample $S_i, i=1,\ldots,m$ by E_0 group on the numerical axis.

This type of problem is new and is being solved for the first time.

3. Hierarchical method of grouping symptoms based on stasis of objects

symbols $X(k)\subset X(n),\ k\le n$ is calculated as follows.

For the unification of scales, the values of quantitative symptoms are reflected in the interval [0,1] by means of fractional-linear reflection.

E_0 Juravlev metric is used as a measure of proximity between two objects of the $S_b=(x_{b1},\ldots,x_{bn})$ sample $S_a=(x_{a1},\ldots,x_{an})$

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

In order to limit the exhaustive sorting in finding an informative subset of symptoms, for each $S_d \in K_p, d=1,\ldots,m$, $p=1,2$ object $X(k)$ on the set of symptoms (1) by the metric E_0 of objects $\rho(S_{d_i},S_d)\le\rho(S_{d_{i+1}},S_d), i=0,\ldots,m-2$ is of non-decreasing order where the inequality holds

$$S_{d_0},S_{d_1},\ldots,S_{d_{m-1}},S_{d_0}=S_d \qquad (2)$$

sequence is constructed.

Let us denote, u_p^1,u_p^2 respectively, (2) $[c_1,c_2](c_2,c_3]$ be the quantities of class objects in the interval that bisects the sequence. $K_p, p=1,2$ Here $c_1=0, c_2=\rho(S_d,S_{d_h})$ and $c_3=\rho(S_d,S_{d_{m-1}})$, h - (2) sequence number.

In defining $[c_1,c_2]$ an interval limit c_2 , $(c_2,c_3]$ each of the intervals is (1) metric relies on the criterion that it contains only values of distances to objects of one class and $X(k)$ its extreme value on the set of symptoms

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

is calculated through the criterion.

(3) on the interval constructed by the criterion $\lambda_1(p)=\left|\left\{S_{d_i}\in K_p\,|\rho\left(S_d,S_{d_i}\right)\in[c_1,c_2]\right\}\right|$, let $\lambda_2(p)=\left|\left\{S_{d_i}\in K_{3-p}\,|\rho\left(S_d,S_{d_i}\right)\in[c_1,c_2]\right\}\right|$, , $\theta_1(p)=\lambda_1(p)/|K_p|$ and $\theta_2(p)=\lambda_2(p)/|K_{3-p}|$. Then $X(k)$ for the set of symptoms $S_d\in K_p$, $p=1,2$ object stability

$$U\left(S_d,X(k)\right)=\theta_1(p)\left(1-\theta_2(p)\right) \quad (4)$$

is determined by its appearance.

Reminder. (3) values of quantitative (initial and latent) symptoms can be used as givens for the criterion. In this case, c_2 it can be considered as the optimal value of the boundary between the classes according to the indicated symptom. Represents the intermediate value of K_1 the criterion $[0,1]$ and the degree of mixing of symptoms in its classes. K_2

In problems of hierarchical agglomerative grouping, there are currently no clear rules and dimensions of proximity for extracting subsets of symbols, using them for mapping values to the number axis in the description of objects. At this point, it is known from the theory of cluster analysis that for most of the grouping methods, the proximity matrix is the initial data.

Given $S_a,S_b\in E_0$ of objects $\left\{x_i,x_j\right\}\subset X(n)$ distance according to symptoms

$$\beta\left(i,j,S_a,S_b\right)=\begin{cases}\rho\left(S_a,S_b\right),i\neq j,\\0,\ \text{otherwise}\end{cases}$$

by its appearance .

the probability matrix b_{ij}, i.e., $\left\{x_i,x_j\right\}$ the size of its contribution to the selection K_1 and classification K_2 of pairs of symptoms E_0

$$b_{ij}=\begin{cases}\dfrac{\sum_{a=1}^{m}\sum_{b=1}^{m}\alpha(\mathrm{S}_a,\mathrm{S}_b)\beta(\mathrm{i},\mathrm{j},\mathrm{S}_a,\mathrm{S}_b)}{\sum_{p=1}^{2}|K_p|\left(\mathrm{m}-|K_p|\right)},\ i\neq j,\\0,\ \text{otherwise}\end{cases}$$

is defined by , here

$$\alpha\left(S_a,S_b\right)=\begin{cases}0,\ S_a,S_b\in K_i,i=1,2,\\1,\ S_a\in K_i,S_b\in K_{3-i}.\end{cases}$$

Symptoms can be grouped as indicated or identified. The basis for the formation of the group is the subjective opinion of the specialist. As an example, in work [4], in order to reduce the space of symptoms, nominal symptoms are grouped, reflected into one quantitative symptom, and it is added to the quantitative symptoms. In defined grouping, groups and their composition are formed based on some criteria. In the grouping algorithm proposed in the article, a criterion based on object stability relations is used.

The estimate of the stagnation increase for the objects of the sample , $X(k)\subset X(k+1)$ given by the condition $X(k)$ and $X(k+1),1\le k<n$ correspondingly calculated according to the subsets of the set, is calculated as follows: E_0

$$M\left(X\left(k\right),X\left(k+1\right)\right)=\frac{1}{m}\sum_{d=1}^{m}\begin{cases}1,\ U\left(S_d,X\left(k\right)\right)\le\\ \le U\left(S_d,X\left(k+1\right)\right),\\ 0,\ \text{otherwise}\end{cases}$$

Below is an algorithm for grouping symptoms:

Step 1. $l=0,\Omega=\{1,\ldots,n\}$;

Step 2. If $\Omega=\varnothing$ so, go to step 6, otherwise $l=l+1,G_l=\varnothing$;

Step 3. $b_{ij}=\max\limits_{u,v\in\Omega}\{b_{uv}\}_{n\times n}$, $G_l=\{i,j\}$, $\Omega=\Omega\setminus\{i,j\}$;

Step 4. If $\Omega=\varnothing$ so, go to step 6, otherwise $Mon(p)=\max\limits_{p\in\Omega}\left\{M\left(G_l,G_l\cup\{p\}\right)\right\}$;

Step 5. If $Mon(p)\ge 0.5$, $G_l=G_l\cup\{p\}$, $\Omega=\Omega\setminus\{p\}$ go to step 4, otherwise go to step 2;

Step 6. That's it.

several ways to map a group of symptoms onto the number axis . An example as an example , RSA and Fisher methods can be shown [5,6] . These methods have certain advantages and disadvantages. The advantage is their computational simplicity, while the disadvantage is the lack of direct application to the space of various symptoms. For this, it is necessary to move to a new quantitative symptom space by expanding the symptom space. In the Fisher method, there is no guarantee that the symptoms falling into each group conform to the normal distribution law. In many cases, it is necessary to use regularization to obtain a numerical solution. In the RSA method, the solution may not be obtained in cases where the value of eigenvalues is complex or multiple.

In contrast to the methods mentioned above, the method of calculating generalized estimates of objects allows one-value reflection in the space of symptoms of different categories without the listed shortcomings [1].

The generalized assessment of the belongingness of the sample objects to the classes (latent symptom) according to the $R_k(S_a)$ identified $G_k(1\le k\le l)$ group is $S_a=(x_{a1},\ldots,x_{an})\in E_0$, $a=1,\ldots,m$ as follows

$$R_k\left(S_a\right)=\sum_{i\in G_k\cap I}w_i t_i\left(x_{ai}-c_2^i\right)/\left(c_3^i-c_1^i\right)+\sum_{i\in G_k\cap J}\mu_i\left(x_{ai}\right)\qquad(5)$$

is calculated by the formula, where w_i – is the i weight K_1 of the symptom, c_1^i,c_2^i and c_3^i is the interval limits determined by (1) for the symptom $t_i\in\{-1,1\}$, $\mu_i(x_{ai})$ and K_2 the contribution of the gradation of the symptom x_{ai} in separating the classes i . i

Generalized grades

$R_k(S_1),\ldots,R_k(S_m)$

don't do it – of his departure (1) according to criterion b we define the value $w(G_k)$ by

Since the sum of quantitative symptoms depends on random values in the generalized evaluation formula, $G_k\cap I=\varnothing$ monotonicity can be maintained only in the process of formation of , groups, and it $(1\le k\le l)$

$$w\left(G_p\right)\ge w\left(G_q\right),p<q$$

determined by the condition

4. Calculation experiment

A sample representing the mentality of representatives of the local people living in our country for a long time and Southeast Asian peoples living in the last century was used for the

calculation experiment [7]. The selection consists of 100 objects, of which 50 K_1 (representatives of the local people), 50 K_2 (representatives of the people living in this place during the last century) belong to the class, each object has 24 nominal values (answers to questions) characterized by

As a result of conducting a computational experiment, three non-intersecting groups of symptoms were formed. Table 1 shows the values of the generalized estimates in the sequence of groups, composition of groups, and the space of symptoms included in groups according to criterion (1). By finding the aggregated estimates of the group symptom objects, it is possible to move to a new latent symptom space whose size is smaller than the original symptom space. This, in turn, provides an opportunity to find hidden patterns by applying other metric algorithms in the space consisting of new latent symptoms.

Table 1.
Grouping of a set of symptoms

Guru h number (latent symptom sign)	**Group and his composition**	**(1) is the criterion value**
1(L_1)	G_1 ={17,19,24,21}	0.572747
2 (L_2)	G_2 ={4,10,14,5,16,20,15,13, 22,2,9,1,12,11,8,18,6,3}	0.530657
3 (L_3)	G_3 ={7,23}	0.246225

the table below confirm that in the process of formation of groups of symptoms, monotonicity is maintained according to criterion (1), that is, it can be seen that each group of symptoms (latent symptoms) is ranked non-increasingly according to its contribution to the classification of objects into classes. .

A high degree of stability of objects in classes has a positive effect on the increase of compactness of the sample. Table 2 shows the number of items with fixed thresholds in the initial raw symptom space, by symptom in each group, and in the new latent symptom space.

Table 2.
Breakdown of stagnation values into intervals.

No	**The space of symptoms**	**Number of objects within the given limits (%)**		
		[0, 0.3]	**(0.3, 0.6]**	**(0.6, 1.0]**
1.	$X(24)$	45	55	0
2.	G_1	2 5	60	15
3.	G_2	78	22	0
4.	G_3	100	0	0
5.	$L=\{L_1, L_2, L_3\}$	16	40	44

From the obtained results, it can be seen that G_1 the stability of objects by group (line 1) and in the space of latent symptoms (line 5) increased (improved) compared to their values in the initial space (line 1). L This is because G_1 the group consists of the most informative symptoms. L it can be seen that the compactness of the classes is improved due to the large

number of objects falling into the intervals and $[0.6;1]$ stagnation in the space of new latent symptoms .$[0.3;0.6]$

Table 3 below presents the stationarity characteristics for some class representatives in the raw and new symptom spaces.

Table 3.

Objects stagnation values

No	Object number (class)	The space of symptoms				
		$X(24)$	G_1	G_2	G_3	$L=\{L_1,L_2,L_3\}$
1.	8 (K_1)	0 . 4896	0 . 6392	0 . 3496	0 . 24 00	0 . 7176
2.	21 (K_1)	0 . 5304	0 . 6688	0 . 448 0	0 . 24 00	0 . 7176
3.	59 (K_2)	0 . 546 0	0 . 6708	0 . 364 0	0 . 2204	0 . 6992
4.	68 (K_2)	0 . 5084	0 . 624 0	0 . 3784	0 . 1984	0 . 774 0
5.	10 (K_1)	0 . 176 0	0 . 044 0	0 . 252 0	0 . 2584	0.0396
6.	46 (K_1)	0 . 1656	0 . 044 0	0 . 2992	0 . 24 00	0 . 046 0
7.	84 (K_2)	0 . 1368	0 . 0676	0 . 26 00	0 . 276 0	0 . 038 0
8.	87 (K_2)	0 . 1024	0 . 1568	0 . 2024	0 . 14 00	0 . 0544

The analysis of table 3 shows that $X(24)$ - objects with large stagnation values in the raw symptom space G_1 and L improved further in the symptom spaces. This state means that these objects are clear representatives (benchmarks) of their class. On the other hand, it can be seen that the objects G_1 and L spaces with low stagnation values in the space of initial symptoms have deteriorated. Such objects "noisy" objects mean that they are the main claim to be included in the collection .

Summary

A method of forming a new symptom space using hierarchical agglomerative grouping based on the stability values of class objects was developed. It was used to extract hidden patterns from the data base (sampling).

The method can be used in the construction of informational models in subject areas with difficult formalization.

References:

[1]Ignatev N.A. Intellectual analysis of data, classification of nepara ¬metric methods and separation of selected object surfaces. - Tashkent: National University of Uzbekistan. Mirzo Ulugbeka, 2009. – 120 p

[2]Duke W.A. Methodology poiska logicheskikh zakonomernostey v predmetnoy oblasti s nechetkoy sistemologieyy : Na primere kliniko-experimentalnyx issledovaniy : dissertation ... doktor tekhnicheskikh nauk : 05.13.01.- Sankt-Peterburg, 2005.- 309 p.: il. RGB OD, 71 06-5/46.

[3]Kolesnikova S.I. Metody analiza informativnosti raznotipnyx priznakov.- Tomsk: Vestnik Tomskogo gosudarstvennogo universiteta, №1(6), 2009.- P.69-80.
[4]Ignatev N.A., Nurjanov Sh.Yu. _ Vybor parametro regularizatsii dlya povyshenia obobshchayushchey obobshchayushchey fonctionnosti discriminantnyx funktsiy // Uzbekiston Republic Weapon Powers of the academy messages . 2014. No. 1 (14). C. 81–87.
[5]Ayvazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. Practical statistics. - M.: Finance and statistics, 1989. - 608 p.
[6]Duda R., Hart P. Recognition of images and analysis scene. - M.: Mir , 1976. - 512 p.
[7]Eshmuratov Sh.A. Prozrachnost prinyatia reshenia pri sinteze iskusstvennyx neuronnyx setae s minimum configuration: Dis . ... candy. tech science - Tashkent, 2008, 120 p