Mathematical Problems of Computer Science 22, 2001, 64–69.

On the Most Informative Indices Selection Method

Harutyun A. Shahumyan†, Nune G. Baghdasaryan‡, Evgueni A. Haroutunian†

† Institute for Informatics and Automation Problems of NAS RA and YSU
‡ Yerevan State Medical University
E-mail: harutiun@yahoo.com, evhar@ipia.sci.am

Abstract

The ideas and properties of variational and Kullback-Leibler distances are used for reduction of analyzed variables space dimensionality and for selection of the most informative indices. The special computer program was written for realization of suggested methods. It was applied in the statistical analysis of data on gastroenterological diseases.

1 Definition of more informative characters

Let the characters of an object (patient) are random variables $X^{(l)}$ with values $x_r^{(l)}$, $r = \overline{1, R_l}$, $l = \overline{1, L}$. Let $\mathbf{X} = \left(X^{(1)}, X^{(2)}, \ldots, X^{(L)}\right)$ is an observation with characters of different kinds, that is the vector $\mathbf{X}$ includes quantitative, qualitative and classification elements. If $X^{(l)}$ is a quantitative character then $x_r^{(l)}$ is the center of r-th interval of whole diapason of admissible values $x^{(l)}$, if $X^{(l)}$ is a qualitative character then $x_r^{(l)} = r$ is the degree defining the quality of that character, if $X^{(l)}$ is a nominal character then $X_r^{(l)} = r$ is the index of the class which includes the object.

For unification of all characters analysis, we replace each $X^{(l)}$ by a R_l-dimensional $\left(X^{(l,1)}, X^{(l,2)}, \ldots, X^{(l,R_l)}\right)$ vector, where

$$X^{(l,r)} = \begin{cases} 0, & X^{(l)} \neq x_r^{(l)}, \\ 1, & X^{(l)} = x_r^{(l)}. \end{cases}$$

The dimension of observed vector increases from L to $L^* = \sum_{l=1}^{L} R_l$.

In following analyzes we assume that all above described transformations already have been done, we denote L^*-dimensional system of characters by the same $\mathbf{X}$ and it's value by $\mathbf{x}$.

Let the auxiliary variables $Z^{(m)}, m = \overline{1, M}$ can be selected from the list of initial indices or can be defined as their certain function $\varphi \in \Phi$, where Φ is the set of all admissible transformations. The presentation of each vector $\mathbf{X} = (X^{(1)}, X^{(2)}, \cdots, X^{(L)})$ by a M dimensional $(1 \leq M \leq 1 - L)$ vector $\mathbf{Z} = \varphi(\mathbf{X}) = (Z^{(1)}, Z^{(2)}, \cdots, Z^{(M)})$, is conditioned by the following important aims and reasonings:

- visual presentation of initial data,
- simplification of calculations,

• compression of statistical data without essential loss of information, etc.

Let us denote the value of vector $\mathbf{Z}$ by $\mathbf{z}$ and its measure of informativity by $\mathbf{I}_M(\mathbf{Z})$. We define the most informative $\tilde{\mathbf{Z}}$ by the following expression:

$$\tilde{\mathbf{Z}} = \arg\max_{\varphi\in\Phi}\{\mathbf{I}_M(\boldsymbol{\varphi}(\mathbf{X}))\}.$$

Let $\mathbf{P}_k(\mathbf{z})$ is the probability of $\mathbf{Z}$ in the $k = \overline{1,K}$ classes. Denoting the measure of their difference in the classes k_1 and k_2 by $\delta\{\mathbf{P}_{k_1}(\mathbf{z}), \mathbf{P}_{k_2}(\mathbf{z})\}$, $(k_1, k_2 = \overline{1,K})$ we take

$$\mathbf{I}_M(\mathbf{Z}) = \sum_{k_1,k_2=1}^{K} \delta\{\mathbf{P}_{k_1}(\mathbf{z}), \mathbf{P}_{k_2}(\mathbf{z})\}.$$

We use the ideas and properties of variational and Kullback-Leibler distances for calculation of $\delta\{\mathbf{P}_{k_1}(\mathbf{z}), \mathbf{P}_{k_2}(\mathbf{z})\}$.

Let $\mathbf{P}_k^{(l)}$ is the distribution of $x^{(l)}$ in the k-th class: $\mathbf{P}_k^{(l)} = \{\mathbf{P}_k(x^{(l)}), l = \overline{1,L}\}$, $k = \overline{1,K}$. The variational distance of k_1-th and k_2-th classes defined by the character $x^{(l)}$ has the following form [2]:

$$\boldsymbol{\Delta}(\mathbf{P}_{k_1}^{(l)}||\mathbf{P}_{k_2}^{(l)}) = \frac{1}{2}\sum_{x^{(l)}\in\mathcal{X}^{(l)}} |\mathbf{P}_{k_1}(x^{(l)}) - \mathbf{P}_{k_2}(x^{(l)})|, \quad k_1, k_2 = \overline{1,K}, k_1 \neq k_2.$$

As a criterion of informativity we suggest to use the following sum:

$$\boldsymbol{\Delta}(k_1, k_2; L) = \sum_{l=1}^{L} \boldsymbol{\Delta}(\mathbf{P}_{k_1}^{(l)}||\mathbf{P}_{k_2}^{(l)}), \quad k_1, k_2 = \overline{1,K}, k_1 \neq k_2.$$

In practical researches the probability distributions are approximated by the relative frequencies. Denoting the frequency of $x^{(l)}$ in the k-th class by $f_k(x^{(l)})$ we get the following expression:

$$\hat{\boldsymbol{\Delta}}(k_1, k_2; L) = \frac{1}{2}\sum_{l=1}^{L}\sum_{x^{(l)}\in\mathcal{X}^{(l)}} |f_{k_1}(x^{(l)}) - f_{k_2}(x^{(l)})|, \quad k_1, k_2 = \overline{1,K}, k_1 \neq k_2. \tag{1}$$

(1) must be calculated for each fixed $M = \overline{1, L-1}$ and for all possible $\mathbf{l}_M = (l_1, l_2, \cdots, l_M)$, $\mathbf{x}(\mathbf{l}_M) = \left(x^{(l_1)}, x^{(l_2)}, \ldots, x^{(l_M)}\right)$ vectors, where $1 \leq l_m < l_{m+1} \leq L$, $m = \overline{1, M-1}$.

We define the most informative combination $\tilde{\mathbf{l}}_M$ as:

$$\tilde{\mathbf{l}}_M = \arg\max_{\mathbf{l}_M} \min_{(k_1,k_2)} \{\hat{\boldsymbol{\Delta}}(k_1, k_2; \mathbf{l}_M)\}. \tag{2}$$

Dimension M, $1 \leq M \leq L$, of the most informative character satisfying (2) is expedient to choose issuing from the statement and requirements of the problem. It is sensible to select such M that the difference $\hat{\boldsymbol{\Delta}}(k_1, k_2, \tilde{\mathbf{l}}_{M+1}) - \hat{\boldsymbol{\Delta}}(k_1, k_2, \tilde{\mathbf{l}}_M)$ is relatively smaller.

We purpose a similar method of informative characters selection with the use of properties of Kullback-Leibler distance [6]:

$$\mathbf{K}(\mathbf{P}_{k_1}^{(l)}||\mathbf{P}_{k_2}^{(l)}) = \sum_{x^{(l)}\in\mathcal{X}^{(l)}} \mathbf{P}_{k_1}(x^{(l)}) \log\left(\mathbf{P}_{k_1}(x^{(l)})/\mathbf{P}_{k_2}(x^{(l)})\right).$$

Let compute the sum for L characters:

$$\mathbf{K}(k_1, k_2; L) = \sum_{l=1}^{L} \mathbf{K}(\mathbf{P}_{k_1}^{(l)} || \mathbf{P}_{k_2}^{(l)}).$$

We suggest the following expression for the selection of the most informative characters:

$$\mathbf{K}'(k_1, k_2; L) = \frac{1}{2} \left(\mathbf{K}(k_1, k_2; L) + \mathbf{K}(k_2, k_1; L) \right), \quad k_1, k_2 = \overline{1, K},\ k_1 \neq k_2.$$

Replacing the appropriate formulas we obtain:

$$\mathbf{K}'(k_1, k_2; L) = \frac{1}{2} \sum_{l=1}^{L} \sum_{x^{(l)} \in \mathcal{X}^{(l)}} \left(\mathbf{P}_{k_1}(x^{(l)}) - \mathbf{P}_{k_2}(x^{(l)}) \right) \log \frac{\mathbf{P}_{k_1}(x^{(l)})}{\mathbf{P}_{k_2}(x^{(l)})}, \quad k_1, k_2 = \overline{1, K}, k_1 \neq k_2. \tag{3}$$

Criterion (3) used in practical researches has the following form.

$$\hat{\mathbf{K}}'(k_1, k_2; L) = \frac{1}{2} \sum_{l=1}^{L} \sum_{x^{(l)} \in \mathcal{X}^{(l)}} \left(f_{k_1}(x^{(l)}) - f_{k_2}(x^{(l)}) \right) \log \frac{f_{k_1}(x^{(l)})}{f_{k_2}(x^{(l)})}, \quad k_1, k_2 = \overline{1, K}, k_1 \neq k_2.$$

We define the most informative combination $\widetilde{\mathbf{l}}_M$ as:

$$\widetilde{\mathbf{l}}_M = \arg \max_{\mathbf{l}_M} \min_{(k_1, k_2)} \{ \hat{\mathbf{K}}(k_1, k_2; \mathbf{l}_M) \}.$$

If the characters have normal distributions and similar covariation matrixes in k_1 and k_2 classes, distance of Mahalanobis is convenient for definition of the most informative indices.

The described methods are included in the statistical program package "SIFAnalysis" created by us. It allows also to check the homogeneity of numerical sequences, to find change points of time series and etc [10].

2 Selection of the most informative symptoms of gastroenterological diseases

The selection of the most informative characters has a great practical importance in the medicine. Revealing the informative symptoms of disease simplifies the diagnosis process and reduces the cost of medical investigation [5][8].

We have used described methods for definition of the most informative characters of gastroenterological diseases.

The research is done on data, registered for 103 patients at the ages from 2 to 15 years. The following characteristics were recorded for each patient:

$x^{(1)}$ - severity of illness (I - the weakest, II - average, III - the strongest),
$x^{(2)}$ - age (2-15),
$x^{(3)}$ - gender (1 - male, 2 - female),
$x^{(4)}$ - signs of chronic intoxication (0 - absence, 1 - presence),
$x^{(5)}$ - hepatomegalia (cm),
$x^{(6)}$ - gallbladder's symptoms (0 - absence, 1 - presence),
$x^{(7)}$ - quantity of total protein (g/l),
$x^{(8)}$ - quantity of general bilirubin (mkmol/l),
$x^{(9)}$ - quantity of conjugated bilirubin (mkmol/l),

$x^{(10)}$ - quantity of β-lipoproteids (unit),
$x^{(11)}$ - ESR (mm/hour),
$x^{(12)}$ - quantity of leucocyte (units$\times$1000),
$x^{(13)}$ - quantity of haemoglobin (g/l),
$x^{(14)}$ - alterations of liver tissue (0 - absence, 1 - presence),
$x^{(15)}$ - anomaly of gallbladder and gall channels (0 - absence, 1 - presence),
$x^{(16)}$ - splenomegalia (cm),
$x^{(17)}$ - cholestasis (0 - absence, 1 - presence),
$x^{(18)}$ - liver hemodynamics alterations (0 - absence, 1 - presence),
$x^{(19)}$ - AST (mkmol/ml$\times$hour),
$x^{(20)}$ - ALT (mkmol/ml$\times$hour),
$x^{(21)}$ - OKT (mkmol/ml$\times$hour).

According to the medical investigations all children were classified into three gastroenterological diseases: gastroduedenital (GD) , biliary pathological (BP), chronic persistiral hepatic (CPH) [4]. According to the severity of their illness ($x^{(1)}$) GD patients were divided into two groups (I, II), BP and CHP patients: into tree groups (I, II, III).

The large number of characters requires the selection of the most informative ones. We have found expedient to calculate the variational, Kullback-Leibler and Mahalanobis distances for the pairs of groups (I, II) and (II, III). Only the six more informative symptoms are mentioned in the tables.

Tables. Variational ($\mathbf{\Delta}$), *Kullback-Leibler* ($\mathbf{K}$) *and Mahalanobis* ($\mathbf{M}$) *distances of illness severity groups by more informative symptoms of GD, BP and CHP diseases.*

	(I, II)	$\mathbf{\Delta}$	**K**	**M**
G	$x^{(8)}$	0.628	4.092	1.606
D	$x^{(21)}$	0.521	2.952	0.95
	$x^{(6)}$	0.472	2.106	1.361
	$x^{(17)}$	0.403	1.878	0.702
	$x^{(13)}$	0.333	1.572	0.412
	$x^{(7)}$	0.333	1.251	0.149

	(I, II)	$\mathbf{\Delta}$	**K**	**M**	(II, III)	$\mathbf{\Delta}$	**K**	**M**
B	$x^{(19)}$	0.393	1.366	0.612	$x^{(7)}$	0.640	3.485	0.998
P	$x^{(21)}$	0.353	1.169	0.254	$x^{(5)}$	0.640	3.314	0.486
	$x^{(7)}$	0.340	1.033	0.008	$x^{(19)}$	0.635	4.772	7.93
	$x^{(12)}$	0.207	0.373	0.207	$x^{(20)}$	0.585	4.135	12.61
	$x^{(16)}$	0.193	0.438	0.451	$x^{(12)}$	0.54	3.989	3.713
	$x^{(13)}$	0.180	0.346	0.014	$x^{(11)}$	0.54	2.514	1.642

	(I, II)	$\mathbf{\Delta}$	**K**	**M**	(II, III)	$\mathbf{\Delta}$	**K**	**M**
C	$x^{(13)}$	0.382	1.507	0.19	$x^{(9)}$	0.650	3.299	0.69
H	$x^{(6)}$	0.354	1.438	0	$x^{(21)}$	0.625	4.830	3.159
P	$x^{(5)}$	0.286	0.748	0.228	$x^{(7)}$	0.525	2.993	1.045
	$x^{(21)}$	0.268	1.039	1.26	$x^{(10)}$	0.525	2.471	0.738
	$x^{(18)}$	0.264	0.823	0.113	$x^{(5)}$	0.400	1.793	3.77
	$x^{(10)}$	0.254	0.951	1.285	$x^{(19)}$	0.375	1.065	0.275

We select the most informative characters on the base of variational and Kullback-Leibler distances. The values of Mahalanobis distances are not reliable as the supposition on normal distribution is not exact caused by the little number of observations.

The computations where done in the computer system STATISTICA 5.5 by the use of package "SIFAnalysis".

References

[1] Aivazian S. A., Buchstaber V. M., Yenyukov I. S., Meshalkin L. D. "Applied Statistics: Classification and reduction of dimensionality". (in Russian). Finansy i statistika. Moscow. 1989.

[2] Aivazian S. A., Rimashevskaia N. M. "Typology of usage". (in Russian). Nauka. Moscow. 1978.

[3] Denisov. L. E., Ushakova T. I., Volodin V. D. "The possibilities of applying personal computer facilities in processing the cancer register data". (in Russian). MediNet. Moscow. 1995.

[4] De Grote J., Desmet V. J., Gedigk et al. "A classification of chronic hepatitis". Lanset. Vol. 2. pp 626-628. 1968.

[5] Glantz S. A. "Primer of biostatistics". (in Russian). Practica. Moscow. 1999.

[6] Gulber E. V. "Calculation methods of analysis and recognition of pathological processes". (in Russian). Medzine. Leningrad. 1978.

[7] Kullback S. "Information theory and statistics". (in Russian). Nauka. Moscow. 1967.

[8] Kuzma J. W. "Basic Statistics for the health sciences". Mayfield Publishing Company. London. 1998.

[9] Lisenkov A. N. "Mathematical methods of planning of multifactorial medical and biological experiments" (in Russian). Meditsina. Leningrad. 1979.

[10] Shahumyan H. A. Package "SIFAnalysis" on testing sequences homogeneity and computing information distances, Transactions of the Institute for Informatics and Automation Problems of NAS RA and YSU: Mathematical Problems of Computer Science, XXII, Yerevan, 2001.

Առավել տեղեկատու հատկանիշների հայտնաբերման եղանակի մասին

Հ. Ա. Շահումյան, Ն. Գ. Բաղդասարյան, Ե. Ա. Հարությունյան

Ամփոփում

Վիճակագրական հետազոտություններում տեղեկատու փոփոխականների ընտրությունը կրճատում է փորձնական տվյալների հավաքման աշխատանքները և մեծացնում է անհատների դասակարգման արդյունավետությունը:

Մեր նպատակն է դասակարգման և տեղեկատու հատկանիշների հայտնաբերման արդյունավետ եղանակների մշակումը: Սույն հոդվածում ներկայացնում ենք այդպիսի մի քանի եղանակներ, որոնք հենվում են փոփոխակային (վարիացիոն) և Կուլբակի-Լեյբլերի հեռավորությունների գաղափարների և հատկությունների վրա: