TALLINNA POLÜTEHNILISE INSTITUUDI TOIMETISED
PROCEEDINGS OF TALLINN TECHNICHAL UNIVERSITY
(Compiler Writing, Data Processing, Programming Subjects)
ТРУДЫ ТАЛЛИНСКОГО ПОЛИТЕХНИЧЕСКОГО ИНСТИТУТА
(translated from Russian)
No. 482 1980
pp. 43 – 50 UDC 518.5:519.2

L. K. Võhandu

Some Methods to Order Objects and Variables in Data Systems

Abstract

In this paper a theory of monotonic system for data tables is presented.
As a result a new family of data ordering methods is developed.

1. Introduction

Acceptance or rejection of scientific hypothesis is the main issue in classical data analysis methods. After this it remains unclear by which way these hypothesis were established, and what amount of observed information were already been utilized in order to put forward the hypothesis.

In addition, it must be emphasized that the practice of qualitative data investigation advancements are ineffective.

The aim of the current paper is to develop and to control in practice new methods of multidimensional data structure discovery. This new methodology founds the global maximum for so-called monotonic systems [1,2,3]. On conceptual level, monotonic system is that for each element of which such an influence function is defined and such that an increase (decrease) of an element results in monotone increase (decrease) of other elements influences in the system (rigorous definitions may be found in the paper). Standing on these results, current paper introduces a class of new methods for multi-parametric system structure investigation, which are quite effective for statistical data analysis.

2. Monotonic systems of observations

2.1. The problem foundation. Depending on the mode of observed empirical system information representation usually a distinction is made regarding data type tables' object-variable, object-object and variable-variable. All these types of table are the foundation for one or another model of information discovery about objects. In recent time such a point of view on data table became a starting point for model development introduced by theoreticians belonging to "Exploratory Data Analyses" community. Known envoy of this community is American statistician Tuki. In SSSR the main stream of papers in this direction emerges from the Control Problems Institute (Moscow, E.M. Braverman, I.B. Muchnik) and from the Institute of Economics (Novosibirsk, B.G. Mirkin).

Consequently, a natural question become apparent regarding technical apparatus allowing developing a theory of data aggregation on abstract level with minimal supposition about the system under investigation and without any statistical (distributive) assumptions.

It appears to be clear, that in resolving the task of empirical system structure discovery, and for definitions of natural aggregates of objects and variables, without any statistical assumptions, a theory of special monotonic systems, developed by J. E. Mullat [1], happens to be satisfactory.

2.2. The methodology of kernel finding procedure in data analysis. The method proposed in [2] for kernel finding procedure in monotonic system W for purposes of data structure investigation makes an acquisition of following requirements:

I. For each element of the system there has to be defined a function π measuring the significance $\pi(w)$ (the weight) of the element w for the system W as a whole.

II. There must be specific rules for recalculating the significances (the influence functions f [1] upon the weights) of system elements provided that all the influences upon the weights for particular system element perform in given direction.

[1] An Example the like functions f may be found in J.E. Mullat, Appendix I, by E. Ojaveer, J. Mullat and L. Vôhandu, "A Study of Infraspecific Groups of the Baltic East Coast Autumn Herring by two new Methods Based on Cluster Analysis," Estonian Contributions to the International Biological Program 6, Tartu (1975) 28-50, http://www.datalaundering.com/download/herring.pdf . A note added by JM

The requirements imposed by pp. I,II leave to the researcher a lot of freedom in his/her choice for influence functions selection as well as for the rules selection for such an influences in the system. The only condition for such a choice emphasis embraces that f and π must be in concord with each other in a way, for example, that after all the elements w exclusion from the system W the weights of all elements $w \in W$ equalize to zero. [2]

Construction of influence monotonic system for objects. In the example an influence scale is used for monotonic system construction on objects straightforwardly upon the data table.

Given data table X $(i = 1,...,N; j = 1,...,M)$ and the table X transfer to its frequency form, e.g., the frequency value z_{ij} of the variable j replaces the entry x_{ij}, the sum $S_i = \sum_{j=1}^{M} z_{ij}$ defines the object i variance, and the whole system variance is the sum $S = \sum_{i=1}^{N} S_i$.

A decrement on which the sum of squares decrease defines the influence of an excluded object from the system. It might be accounted for, in view of the calculation process organization practicality, that the excluded object been transferred into special vague class.

Since for each variable j the frequency h_j also has hitherto $h_j - 1$ to itself equal values the sum of frequency squares for the whole system decreases by value

$$g_j = (h_j - 1) \cdot (h_j^2 - (h_j - 1)^2) = 2 \cdot h_j^2 - 3 \cdot h_j + 1.$$

[2] According to the information available, the up to date idea, which actually highlight these requirements on higher level of abstractions leads to data sets organized in so-called antimatroids data set systems, see "Correspondence between two antimatroid algorithmic characterizations, " Yulia Kempner and Vadim E. Levit, Department of Computer Science, Holon Academic Institute of Technology, 52 Golomb Str., P.O. Box 305, Holon 58102, ISRAEL, yuliak,levitv@hait.ac.il. JM, http://www.datalaundering.com/download/0307013.pdf .

The next step will be a function construction, which defines the value of changes in influences of other objects upon the system when one object is excluded. An exclusion of an object k from the system changes the influence of every other object i by the value

$$S_i(h_{ij}) - S_i(h_{ij}-1) = \sum_{j=1}^{M} \delta\left[3\cdot h_{ij} - 2\cdot h_{ij}^2 - 1 - 3\cdot(h_{ij}-1) + 2\cdot(h_{ij}-1)^2 + 1\right] = \sum_{j=1}^{M} \delta(-4\cdot h_{ij} + 5),$$

where $\begin{cases} \delta = 1, \text{if } x_{kj} = x_{ij} \\ \delta = 0, \text{if } x_{kj} \neq x_{ij}. \end{cases}$

Let investigate the formula's monotonicity, which defines the value of changes in influences. Supposing that $M = 1$ and $h_{ij} = 1$ the change is equal to $+1$. In case when $M = 1$, but $h_{ij} \geq 2$, the change is negative. Thus, excluding the object, this ordinary function is not monotonic (so called $\ominus$-action, according to [1], take place towards an element in the object-variable system). However, the weakness can be removed: instead of influence function $2\cdot h_j^2 - 3\cdot h_j + 1$ simple addition of *one numbers* in all histograms classes for variable j results in function

$$g_j = (h_j - 2)\cdot\left(h_j^2 - (h_j - 1)^2\right) = (h_j - 2)\cdot(2\cdot h_j - 1) = 2\cdot h_j^2 - 5\cdot h_j + 2.$$

Now the exclusion of an abject k from the system changes the influence of every object i by the value

$$S_i(h_{ij}) - S_i(h_{ij}-1) = \sum_{j=1}^{M} \delta\left[5\cdot h_{ij} - 2\cdot h_{ij}^2 - 2 - 5\cdot(h_{ij}-1) + 2\cdot(h_{ij}-1)^2 + 2\right] = \sum_{j=1}^{M} \delta(-4\cdot h_{ij} + 7).$$

Since the least real frequency $h_{ij} = 2$, it is easy to verify that after the *ones addition* to all histogram classes, the change in influences is monotonic.

2.3. The influence function in two-dimensional case. In previous section we introduced an influence function f, which guaranteed the monotonicity in case of $\ominus$-operation upon system objects. Thus, the function definition range represents a set of objects. Usually, after the extraction of objects group a natural question emerges regarding the variables by which these objects constitute a separate class. Similar interpretation problem emerges in all classification methods wherein we utilize a matrix of distances between the objects. After the group has been found an interpretation of the results is necessary since it is not clear what variables lie in the foundation of the group.

Therefore, it is important also to develop methods allowing concurrent partition of objects and variables lessening herein the result interpretation.

A variety of variants for π function definition are available in the system of objects. Two main categories of influence function make a distinction – the additive and the multiplicative. In principle, for each category there exists an infinite number of actual influence functions.

2.4. Additive category of influence functions. An additive system on frequency data tables of weight functions are called functions type

$$S_i = \sum_j g(n_{ij}),\ P_j = \sum_i g(n_{ij}),$$

where $i = 1,...,N$; $j = 1,...,M$, $g(x)$ is a function of frequency.

In order to guarantee the kernel splitting method realization capability it is necessary to stipulate that function $g(x)$ value within the range of actual frequencies x is non-decreasing while shifting from $g(x)$ to $g(x+1)$ (or non-increasing while shifting from $g(x)$ to $g(x-1)$).

As a consequence, in the position of weight function one can formally use a lot of functions common to mathematical analysis course. However, the demand of

conceptual interpretation of the classification results brings in force only weight functions, which have an appropriate interpretation capability. Different entropy functions and the like belong to such functions.

Instead of the classical entropy set X function $H(X)$ it is better to use its approximation

$$H_2(X) = \sum_i p(x_j)^2 .$$

Nilson [4] showed that the entropy $H_2(X)$ defined by the method above preserves all the properties of ordinary entropy and may provide a basis for rich multidimensional statistics. For our purposes, it is important that for vector pair X and Y one can pull together an ordinary two-dimensional table with frequencies n_{ij}, $i = 1,...,p$; $j = 1,...,q$ whereas the mutual information (see [4]) yields to

$$I(X,Y) = N^2 \cdot \sum_i \sum_j n_{ij}^2 \cdot \left(\sum n_{i\bullet}^2 \cdot \sum n_{\bullet j}^2 \right)^{-1} .$$

Yet another type of additive weight functions belongs to the category of objects influence function $g(n)$ on the system (see, above, 2.4.).

The kernels finding procedure, KFP, executes only after the weight function have been chosen and the monotonicity confirmed by the direction of changes in the weights as a result of $\ominus$-actions upon the system elements. Together with this, it is important to establish the strategy of KFP procedure implementation.

For example, following strategies in the system object-variable are possible, which guarantee the KFP implementation.

I. The KFP implementation originates in objects elimination starting from the object with the minimal weight when the role of the system element w been assigned to the object.

II. By the rules of KFP it organizes the elimination both the objects and variables starting from the element with minimum influence on the system when system element w role been assigned either to the object or to the variable.

III. When the role of an element w been assigned to the entry in data table belonging to the row i in the column j then by the KFP regulations the elements elimination starts from:

a) $\min\limits_{i,j} \pi(R_i, V_j)$,

b) $\min\limits_{i,j} (R_i + V_j)$,

c) $\min\limits_{i,j} (g_{ij} + \bar{g}_{ij})$,

where g_{ij} is the influence function of the variables histogram;

$\bar{g}_{ij}$ is the same influence function but this time of objects histogram.

2.5. Multiplicative category of influence functions. We suggest together with additive weight functions the following multiplicative type of weight functions:

a) define the weight of a data table $X = \|x_{ij}\|$ element as a number

$$w_{ij} = g_{ij}(x_{ij}) \cdot \bar{g}_{ij}(x_{ij}),$$

where the function value $g_{ij}(x_{ij})$ represents a frequency n_{ij} of a data table element x_{ij} upon the variable j;

the same value $\bar{g}_{ij}(x_{ij})$ but calculated upon the frequencies histogram of object i.

An example of function $g(x)$ arrives by function $g(n) = 2 \cdot n^2 - 5 \cdot n + 2$ or by function $g(n) = n^{\beta} - 1$ ($\beta \geq 1$);

b) define the data table element's x_{ij} weight as a number

$$w_{ij} = \sum_i g_{ij}(x_{ij}) \cdot \sum_j \bar{g}_{ij}(x_{ij}).$$

The strategy may vary how to select the monotonic system element:

a) select an element w as an object, then we eliminate the object totally; [3]

b) select an element w as an object or variable;

c) select an element w as a data table element defined by its indices i and j.

3. Monotonic system kernel splitting algorithm on data table

We expose the algorithm in steps. Consider a two-dimensional $N \times M$ array X, whose elements are the natural numbers within the range from 0 to 255.

A0. Using the data table X calculate table of frequencies stripes z (the histogram table Z) for variables. Initiate arrays S and V: $S(I)=0$ for $I=1,...,N$ and $V(J)=0$ for $J=1,...,M$.

A1. (The influences calculation). For $i=1(1)N$, $j=1(1)M$ calculate $g_{ij}=5\cdot z_{ij}-2\cdot z_{ij}^2-2$; $S(i)=S(i)+g_{ij}$; $V(j)=V(j)+g_{ij}$.

A2. (To eliminate the object or the variable?). Find $h=\max\limits_{i} S(i)$, $g=\max\limits_{j} V(j)$ and $max(h,g)$. Go to A3 or A4 correspondingly.

A3. (The variable elimination). For all not yet eliminated columns-variables recalculate $S(i)$: $S(i)=S(i)-g_{ij}$. Go to A5.

A4. (The object elimination). For all not yet eliminated rows-objects recalculate $S(i)$: $S(i)=S(i)+P_i$, where $P_i=4\cdot T-7\cdot L$ ($T-$ matching variables frequencies sum in object i) and find $h=\arg\max\limits_{i} S(i)$ while L is the number of frequencies matches; together with the numbers P_i calculation we adjust as well $V(j)=V(j)+(6\cdot z_i-20)\cdot z_i+16$.

In the histogram table Z for all not yet eliminated values of object h subtract 1. Equalize $S(h)=\infty$ in order to indicate that the object h has been eliminated.

A5. In case there are not yet eliminated objects or variables go to A2, otherwise finish.

[3] By the rules of the KFP procedure we eliminate not only the minimal weight object, but also all objects in the elimination steps, one by one, recalculating frequency histograms after the elimination take place within each step. Noted by JM

The data table X is send to the printer after the algorithm has finished, but first with permutated rows and columns in the order objects or variables has been eliminated. The kernel is placed into the right-down corner towards the print layout. Henceforward we move along the sequence of the monotonic system elements till the first local maximum is found among the weights in the moment the elimination happens. All the elements of the sequence in backward direction till the local maximum inclusive belong to the kernel.

Literature

1. J.E. Mullat, "Extremal Subsystems of Monotonic Systems, I," *Automation and Remote Control, 1976, 37*, 758-766, http://www.datalaundering.com/download/extrem01.pdf .
2. J.E. Mullat, "Extremal Subsystems of Monotonic Systems, II" *Automation and Remote Control, 1976, 37*, 1286-1294, http://www.datalaundering.com/download/extrem02.pdf .
3. J.E. Mullat, "Extremal Subsystems of Monotonic Systems, III," *Automation and Remote Control, 1977, 38*, 89-96, http://www.datalaundering.com/download/extrem03.pdf .
4. Nilson A. "The quadratic sums properties," *Proc. of Estonian Academy of Science, 1965*, Ser. Physic-Math. Sciences.