CLUSTER ANALYSIS OF SOCIO-ECONOMICAL DATA

N.Y. Blagoveschenskiy, Scientific Center of Legal Information, nikblag@scli.ru

A.A. Rubchinsky, NRU HSE and University «Dubna», arubchinsky@yahoo.com

1. Introduction

In investigation of systems, in which human factor is the determining one, it is doubtful to obtain adequate enough reality presentation in form of traditional (or so called hard) mathematic-cal models. Therefore cluster analysis, i.e. selection of several groups of similar in some sense objects from the whole considered collection, is one of the most fitting research tools in such ill formulized cases. In classification problems it is required additionally that the selected clusters form a division of the initial set, but the abandoning of this requirement seems more realistic in the considered situations. Informal character of the clustering problem, its various modifica-tions, statements and applications, numerous approaches and methods of its solution are compre-hensively described in several monographs and reviews (see, for instance, Braverman and Much-nik, 1983, Filippone et all, 2008, Gordon, 1999, Luxburg, 2007, Mirkin, 1996, Mirkin, 2005, Mirkin, 2010).

Our main concern is focused on formalization, exact definition and calculation of the im-portant property of subsets of the given initial set that describes their stability, exactness, validity – in essence, their possibility (or impossibility) to be selected as clusters. This property is named ***volatility***, which is determined formally for separate candidates as well as for the whole clustering problem.

Let us consider some examples without giving exact definition of volatility but rather to give some hints to its future definition. Clusters with different volatility are shown in Fig.1. In Fig.1*a* and 1*b* three considered clusters have volatility 0, despite the fact that selection of clusters in Fig.1*b* is more difficult than selection of the same clusters in Fig.1*a*. The clusters shown in Fig.1*c* have different volatilities. Intuitively cluster 1 has the same volatility 0, cluster 2 has some small volatility, and volatility of cluster 3 exceeds volatility of cluster 2. Finally the cluster 3 in Fig.1*d* practically disappears (its volatility is close to the maximal number 1) , meanwhile clusters 1 in all the pictures has the same volatility, as well as cluster 2 in Fig.1*c* and 1*d*.

Usually the notions of volatility, stability, and so on are connected with the process of changes of a considered system in dependence on time or other external parameters. In the sug-gested approach to clustering, however, this it is not the case. Volatility is determined for a given clustering problem. The essence is that the suggested clustering algorithm (like some other ones) consists of repeating randomized steps. At every step a family of subsets (candidates for clusters) is constructed. Clear-cut clusters with zero volatility are absolutely the same at every algorithm run. Less clear clusters can be slightly different or / and occur not at every algorithm run. This reasonning enables to formulate a simple formal criterion, whose maximization defines volatility of a given cluster. The volatility of the whole clustering problem is determined as weighted sum of volatilities of the found clusters.



Fig.1. Clusters with different volatility

The idea of duality of system dynamics and statics (sometimes named as idea of canonic ensemble) is one of the most essential ideas of natural sciences. This idea is especially important near phase’s transitions, bifurcation points, etc. However, in investigation of socio-economical systems this approach is comparatively uncommon. One of the few exceptions is the book (Weidlich, 2000).

It seems that high level of volatility corresponds to difficulty of a clustering problem, and realization of this connection led to the new clustering algorithm. This algorithm finds the clusters with arbitrary levels of volatility (including the conventional case of zero volatility) that enables to cope with hard clustering problem as well as with easy ones. Moreover, the feasible level of volatility is one of very few external parameters of the suggested algorithm. It is practically one that is essentially depends upon human decision.

The goal of the article consists in presentation of the new clustering algorithm satisfying the following requirements:

* clustering results do not contradict to intuitively clear answers in various simple situations;
* no assumptions of stochastic, geometric, and other characters are made;
* number of clusters is determined only in the process of the algorithm running, particularly, the absence of clusters is possible;
* the algorithm uses very few parameters with clear interpretations;
* human decision (if necessary) is made only at final stage.

The considered formal presentation of clustering problems and the structure of the suggested clustering algorithm is described in Section 2. The formal definition of volatility and the new algorithm of clusters selection based on their volatility are given in Section 3. The model and real examples are considered in Section 4.

2. The structure of the clustering algorithm

In the suggested approach initial data about the problem are presented by the well-known neighborhood graph (see, for instance, Luxburg, 2007). Graph vertices are in one-to-one correspondence to given objects. Any vertex *v* is connected to 4-5 other vertices, corresponding to objects the most close to the object corresponding to vertex *v*. The proximity of objects is determined by an initial problem description: a given similarity **⁄** dissimilarity matrix, pattern matrix (objects **⁄** parameters) or by many other types of description. Advantages and disadvantages of neighborhood graph presentation are not discussed here. The most important – and only essential – justification of any clustering method consists in its good fits with experimental results and common sense. However, this presentation deals with the most «soft» data about connections between objects that are classified. In the framework of the suggested approach only these – essentially qualitative – data about connections are used.

The algorithm is determined as a three-level procedure. The internal level of the suggested procedure consists in the dichotomy of any undirected graph. The intermediate level produces a family of classifications of the initial set of objects. All classes in all these classifications are unions of the same *K* «bricks» – classes, received as results of *K*–1 consecutive dichotomies by the algorithm from internal level. At every step a subset with the maximal number of element (among all the found by this step) is selected for the next division. This algorithm was named as Divisive-Agglomerative Classification Algorithm (DACA for brevity). Finally, the external level deals with families of classifications, constructed by several runs of the intermediate level algorithm. The internal and intermediate levels are comprehensively described in Rubchinsky, 2010. Therefore they are not discussed here. Instead, in the next Section we focus our attention at the external level of the suggested new three-level procedure.

3. Cluster Selection Algorithm.

Before starting the algorithm description, let us describe its input in more detail. Assume *r* is the number of independent runs of DACA. Because every run uses random numbers (for ins-tance, for consecutive choice of pair of vertices in the internal minimax algorithm), DACA pro-duces at every run a family of classifications. Generally speaking, these families can be different, though they coincide in many simple cases. Moreover, the quantitative measure of their coinci-dence (that will be defined in this section) can be considered as a formal measure of complexity of a given clustering problem.

Let us introduce some necessary definitions and notations. Assume *Ui* is the set of all the classes included in all the classifications found by DACA at its *i*-th run. All the elements of *Ui* (*i* = 1, …, *r*) are candidates for clusters. For simplicity, they are named «clusters».

Assume F be an arbitrary family of clusters, belonging to different sets *Ui*. It will be convenient to present F as follows:

F = 〈$F\_{i\_{1}}, …, F\_{i\_{d}}$〉, where $F\_{i\_{k}}$∈ $U\_{i\_{k}}$ (*k* = 1, …, *d*}, and *s* < *t* implies *is* < *it*. (1)

Denote

*A*(F) = $∩$*Fj*, *B*(F) = $∪$*Fj* , *α*(F) = **|***A*(F)**|** ⁄ **|***B*(F)**|**, (2)

If *α*(F) >0.5, then this family of cluster is named *α-****stable***. The number *α*(F) gives some presenta-tion about the stability of an arbitrary family of clusters. Results of 4 runs of DACA are shown in Fig.2. The three families of clusters P, Q and R are separately shown in Fig.3 in more detail.



Fig.2. Results of 4 runs

At the same time Fig.2 and 3 point out to the other notion of stability. It is worthwhile to take into account that clusters from family P appear 3 times of 4, clusters from family Q appear 4

****

****

****

Fig.3. Family intersection and union

times of 4, and clusters from family R appear 2 times of 4. Denote *c*(F) = **|**F**| =** *d* (see (1)). As-sume

*β*(F) = *c*(F) **⁄** *r*. (3)

This parameter shows, how many times of *r* runs components of F are included in found families.

The introduced values *α*(F) and *β*(F) are the reverses of each other: addition of a new cluster in the arbitrary family F increases *β*(F) but decreases *α*(F), while deletion of any cluster from the arbitrary family F decreases *β*(F) but increases *α*(F). Therefore the product

*γ*(F) = *α*(F) × *β*(F) (4)

is considered as the measure of stability of family F, and the supplementary value

*V*(F) = 1 – *γ*(F) (5)

is named the ***volatility of the family*** F. Of course, many analogous expressions for stability (and, hence, volatility) of family can be written (for instance, *γ*(F) = 0.5(*α*(F) + *β*(F))), but these forms do not essentially affect the clustering results. Define the volatility

Assume *Ai* = *A*(F*i*), *i* = 1, 2, … (see formula (2)). These sets are black figures in the middle column in Fig.3. Finally, assume the number *V*\* – the maximal feasible volatility – is given.

The following steps of the **Algorithm of Clusters Selection** define the suggested solution of a given clustering problem. The input of the algorithm consists of *r* family of classifications, found at the intermediate level (see Section 2).

**Algorithm of Clusters Selection**

1. Find all the ***α-stable***  families F1, F2, …, F*m* (see (2)).

2. Select among them all the families F such that *V*(F) ≤ *V*\* (they are named the ***feasible*** ones).

3. Order feasible families F*i* in correspondence with *V*(F*i*) increasingly.

4. Define sets *Ci* = *A*(F*i*) (*i* = 1, 2, …, *k*).

5. Assume *D*1 = *C*1, current *ic* = 1.

6. If sets *D*1 , …, *Dt* are found, consider consecutively *i* > *ic* till one of the following two events occur:

* *Ci* does not intersect with *D*1 , …, *Dt*;
* *i* = *k*+1.

In the 1-st case assume *Dt* +1 = *Ci*, *ic* = *i*, *t* = *t*+1and return to step 6.

7. Consider all the clusters *D*1 , …, *Dt* and eliminate every cluster containing other clusters from the list.

8. Stop.

The constructed sets *D*1 , …, *Ds* form the output of the external stage 3. Sets *D*1 , …, *Ds*are the found clusters. The ***volatility V(D) of cluster D*** is defined as volatility of family F such that *D* = *A*(F). Thus, volatilities of the constructed clusters *D*1 , …, *Ds* are ordered increasingly. The volatility of the whole clustering problem is defined as the weighted sum of all the found clusters:

*V* = $\sum\_{i=1}^{s}V(D\_{i})|D\_{i}|$ **⁄**  $\sum\_{i=1}^{s}|D\_{i}|$. (6)

In order to resume this Section, let us describe the operation of Step 1 – **Construction of** α**-stable Families.** The algorithm is rather simple. We construct the list of all families F, such that *α*(F) > 0.5. Assume we have already the current list of such different families F1, …, F*s*. Assume F = 〈$F\_{i\_{1}}, …, F\_{i\_{d}}$〉 is one of constructed families, presented in form (1). Consider arbitrary set *Fi* from any set *Ui*, where *i* > *id*. Check the new family F*’ =* 〈$F\_{i\_{1}}, …, F\_{i\_{d}}$, *F*〉 for the condition *α*(F’) > 0.5 (it is a simple operation). If this condition holds, F*’* is added to the list.

The same operations is executed

1) for all the elements of *Ui*;

2) for all *i* (*id* < *i* ≤ *r*);

3) for all the families of the current list.

The algorithm stops then no new family cannot be added to the current list. Initially all the separate sets from every *Ui* (*i* = 1, …, *r*) form the current list ■

It is worthwhile to remark that the algorithm is fast enough, because for almost all pairs of two sets from different *Ui* their intersection is empty and therefore all the chains $F\_{i\_{1}}, …, F\_{i\_{d}}$ are very quickly terminated.

4. Real Examples

1. Stock market analysis. The results are presented in terms of the found groups of stock for USA, Russia and Sweden stock market. In the considered case the initial data consist of pair-wise correlations for 2008-2010 years: 500 USA stock; 266 Sweden stock; 151 Russian stock. It is required to find clusters in these data (or be sure in their absence).

USA market. Volatilities of the found clusters are presented in the following table:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| №№ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Volatility | 0.000 | 0.125 | 0.167 | 0.167 | 0.286 | 0.356 | 0.549 | 0.583 |

It is possible to add that all the clusters are formed by companies engaged in the same or close fields. The cluster with the minimal volatility 0 includes only the companies engaged in the same field (gold mining). The increasing of volatility is accompanied (as a trend) by the widening of field of activity of included firms. The obtained clustering results at least do not contradict to common sense.

Russia market. Only 2 clusters are revealed in Russian stock market. Both groups of companies are engaged in electrical power production. In both cases volatility is equal to 0.15. One group consists of 18 companies, the other consists of 5 companies. The correlation between stock, included in the same cluster, is significantly less than in USA market. This circumstance demonstrates the significant difference between these two markets.

Sweden market. Under the same algorithm no clusters in Sweden data are revealed.

2. Deputies Clusters in Duma. In this case the activity of Russian Duma (parliament) was analyzed for period of 5 months, since 01.09.2001 till 01.02.2002. This period seems important, because the significant political event –occurrence of new party «Unified Russia» – happened 01.12.2001. In more detail the situation is described in book Aleskerov et al, 2006. Five families of classifications (corresponding to the considered five months period) are found. For every separate month all the votings (200 – 500) are considered. To *i*-th deputy (*i* = 1, 2, …, 479) a vector *vi* = ($v\_{1}^{i}$, $v\_{2}^{i}$, …, $v\_{n}^{i}$) is related, where *n* is the number of votings in the months,

$v\_{j}^{i}$ = $\left\{\begin{matrix}1, if i-th deputy voted for j-th proposition; \\-1, if i-th deputy voted against j-th proposition; \\ 0, otherwise. \end{matrix}\right.$

The dissimilarity *dst* between *s*-th and *t*-th deputies is defined as usual Euclidian distance between vectors *vs* и *vt*. The dissimilarity matrix *D* = (*dst*) is the initial one for clustering algorithm, described in Section 2. The volatilities of all the clusters are presented in the following table:

Volatility in duma clusters

|  |  |
| --- | --- |
| September |  0; 0; 0; 0; 0  |
| October |  0; 0; 0; 0; 0; 0; 0; 0 |
| November |  0; 0; 0.022; 0.200; 0.260; 0.315 |
| December |  0; 0; 0; 0.010; 0.012; 0.074; 0.125 |
| January |  0; 0; 0; 0.020; 0.035; 0.060; 0.144 |

This table includes only the most stable clusters with *α*(F) > 0.8 (see (2)). Of course, it seems important to know the party whose fractions form the clusters with minimal and maximal volatility. The answers are the following:

1. Only one fraction – Unity (Единство) – forms the clusters with 0 volatility for all the 5 months.
2. The fraction OVR (Fatherland is all the Russia) does not form its own cluster; more-over, its members are included in different fractions with high volatility levels.
3. The volatility of the whole problem (see (6)) was equal 0 in September and October 2001; it significantly increased just before the key event – the creation of the party «United Russia» as a result of joint of two parties: Unity and OVR, and slightly decreased after this event.

In the cited book Aleskerov et al, 2006 several known indices of Duma did not show sig-nifycant features at this period.

The authors are partially financially supported by LATNA Laboratory, NRU HSE, RF government grant, ag. 11.G34.31.0057.

References

Aleskerov F.T et al. Влияние и структурная устойчивость в Российском парламенте (1905–1917 и 1993–2005 гг.). – М.: ФИЗМАТЛИТ, 2007. – 312 с.

Braverman E.M., Muchnik I.B., Structured Methods of Experimental Data Processing (in Russian) – Moscow: Nauka, 1983.

Filippone M., Camastra F., Masulli F. Rovetta S. A Survey of Kernel and Spectral Methods for Clustering. – Pattern Recognition, 41(1):176-190, January 2008.

Gordon, A.D. Classification. – Chapman & Hall/CRC, 1999..

Luxburg, U. A Tutorial on Spectral Clustering. – Statistics and Computing 17(4): 395-416, 2007.

Mirkin, B. Mathematical Classification and Clustering. – Kluwer Academic Publishers, 1996.

Mirkin, B. Clustering for Data Mining: A Data Recovery Approach. – Chapman & Hall/CRC, 2005.

Mirkin, B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. – Sprin-ger, 2010.

Rubchinsky A. Divisive-Agglomerative Classification Algorithm Based on the Minimax Modifi-cation of Frequency Approach: NRU HSE, Moscow. Working paper WP7/2010/07. – 48 p.

Weidlich W. Sociodynamics. A Systematic Approach to Mathematical Modelling in the Social Sciences. – Harwood Academic Publishers, 2000 (Russian Translation: Вайдлих В. Социодинамика: системный подход к математическому моделированию в социальных науках. М.: – Едиториал УРСС, 2005.)