Overlapping Clustering of Binary Variables

Dušan Húsek
ICS, Academy of the Science of the Czech Republic
Pod vodárenskou věží 2, 182 07 Praha 8, Czech Republic
e-mail: dusan@cs.cas.cz

Hana Řezanková
Dept. of Statistics, University of Economics, Prague
nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
e-mail: rezanka@vse.cz

Alexandr A. Frolov
IHNA, Russian Academy of the Sciences
Buttlerova 5a, Moscow, Russia
e-mail: aafrolov@mail.ru

Keywords: machine learning, knowledge extraction, overlapping clustering, clustering of variables, fuzzy cluster analysis, factor analysis, neural networks

1. Introduction

The process that groups similar objects together and forms clusters is often called segmentation or clustering. Clustered groups are homogeneous within and desirably heterogeneous in between. The rationale of intra-group homogeneity is that objects with similar attributes are likely to respond in a similar manner to a given action. This property has various uses both in business and in scientific research. In addition to cluster analyses, we can use multidimensional scaling, factor analysis or correspondence analysis for graphical representation of clusters

The very often problem of clustering is separability of clusters because of their overlapping nature. To solve this problem some special techniques were developed for overlapping clustering, see (Gordon, 1999). Another way to cope with this problem is fuzzy cluster analysis. As result of this analysis we obtain membership values for each of the variables and each cluster. These values identify that the variable can be assigned to more than one cluster and level of the assignment.

The specific problem is analysis of the binary data which demands either special similarity measures or special methods. As regards last mentioned only monothetic analysis (Kaufman and Rousseeuw, 2005) come to commercial implementation.. However, this method is not suitable for finding overlapping clusters. On the other side, implementation of fuzzy cluster analysis that is suitable for overlapping data, does not allow using special measures foe binary data. To summarize, commercial products do not support straightforward way for overlapping clustering of binary data (OCBD).

In this contribution we propose some possible solutions for OCBD. The first approach stresses on combinations of standard statistical procedures. We suggest applying factor analyses or multiple correspondence analyses and interpreting the factor loading matrix (or coordinates) by fuzzy cluster analysis. This technique leads to an easier identification of clusters. In the second approach we propose to use the nonlinear binary factor analysis

algorithm based on Hopfield-like neural network. Main advantage of this method should be applicability on very large data sets. This method is the result of our research.

2. Combination of commercially implemented statistical methods

One way of finding overlapping clusters is by using a combination of the results of different clustering techniques. Variables, which belong to only one cluster, are mostly identified identically when using different techniques. Variables, which belong to two or more clusters, can be assigned differently when using different methods. The disadvantage of this approach is its complexity.

The other way to obtain overlapping clusters of variables is the application of factor analyses or multiple correspondence analyses and interpretation of the factor loading matrix (or coordinates) by fuzzy cluster analyses. By this way we can better differentiate variables which belong only to one cluster and variables which can be assign to more than one cluster.

We can suppose that the number of clusters (L) is known. Firstly, we need to determine the number of factors (K) or coordinates, respectively. We can suggest the K value so that the following relations were satisfied:

$$K_{\min} = 2, \quad 2^{K_{\max}-1} < L \quad \text{and} \quad 2^{K_{\max}} \geq L$$

From the fuzzy cluster analysis, we obtain memberships u_{iv} for each of the variables $\boldsymbol{X}_i$ and each cluster v. Memberships have to satisfy the following conditions:

1. $u_{iv} \geq 0$ for all $i = 1, 2, \ldots, M$ (M is a number of variables) and all $v = 1, 2, \ldots, L$,
2. $\sum_{v=1}^{L} u_{iv} = 1$ for all $i = 1, 2, \ldots, M$.

For applications, we used the fuzzy cluster analysis implemented in the S-PLUS statistical package. The memberships are defined through minimization of function:

$$f = \sum_{v=1}^{L} \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} u_{iv}^2 u_{jv}^2 d_{ij}}{2\sum_{j=1}^{M} u_{jv}^2}$$

where dissimilarities d_{ij} (dissimilarity between i^{th} and j^{th} variables) are known and memberships u_{iv} and u_{jv} are unknown.

We suggest interpreting the results in the following way: If the value of a membership coefficient is greater than or equal to $1/L$ then the variable belongs to the particular cluster (this value should be replaced by 1); in the opposite case the variable is not assigned to this cluster (this value should be replaced by 0).

3. Hopfield-like neural network

This approach is based on the idea of factor analysis when the input data matrix can be expressed by the multiplication of two matrices. One of them contains factor scores and the second contains factor loadings. A developed neural network is capable of analyzing binary data. We can also obtain both matrices with binary data.

As shown in paper (Frolov, 2004), a modified version of the binary Hopfield neural network can be used to find binary factors in terms of Nonlinear Boolean factor analysis. We proposed a method for a sufficiently nonlinear case that naturally follows from Hebbian-like

learning and properties of Hopfield-like neural network dynamics. This conceptual mechanism seems plausible for memory storage and processing in the brain. This algorithm has several additional requirements on analyzed data (which must be fulfilled in order let it work correctly), but is able to work with larger data files.

We suppose that each i^{th} case (row) $\mathbf{X}_i$ (i = 1, 2,…, N) of the analyzed data matrix can be expressed as a logical sum of weighted vectors of factor loadings:

$$\mathbf{X}_i = \bigvee_{l=1}^{L} f_{il}\boldsymbol{\alpha}_l$$

where f_{il} are factor scores, $\boldsymbol{\alpha}_l$ are vectors of factor loadings (l = 1, 2,…, L) and L is a number of factors.

The principal requirement on the developed method was that factors are supposed to be found sufficiently quickly even if we have no any prior information about them and much faster if we do have such information. Starting from a random initial state, the network activity stabilizes in some attractor which corresponds to one of the factors or spurious states.

This convergence is very fast: the strongest factors could be revealed for only several iteration steps. To separate true and spurious attractors we found a procedure based on calculation of their Lyapunov function (Husek, Frolov at al., 2005). The highest the Lyapunov function is the strongest the factor is. The unlearning of already found factors prevents their repeated retrieval. Some background on this topic can be found in work (Goles-Chacc, Fogelman-Soulie, 1985).

4. Applications

We tested our procedure over different examples from literature and test collections. Here we applied techniques mentioned above to two the different data sets. First we analyzed data matrix representing a Reuters collection of text documents and found around 80 topics. And as a second benchmark we used Russian parliament voting data, using the results of roll-call votes in the Russian parliament during 2004.

4.1 Reuters collection analysis.

Nonlinear Boolean factor analysis of text documents implies that we are looking for groups of highly correlated words each of which represent one topic, and then describe documents in terms of these revealed group of words – topics. So by means of Boolean factor analysis someone could mine from document collection its common properties and general features, construct rubrics and reduce dimension of documents representation.

We applied our method to the set of 21000 messages of agency Reuters (Reuters, 2004) as well. Each message was transformed to binary code dependently on presence or absence of words in the message. The used vocabulary contained 5000 the most often words in the set (consequently network contained 5000 neurons).

It can be seen that our method combines words in factors not only according to the frequency of their appearance together at the messages but mainly according to their appearance at the same context. We see that different factors reflect different contexts of word utilization and different topics of news messages, while messages with the same topics are connected with the same factors.

Two messages with highlighted words creating factors are shown (Tab. I. and II.), as an

example of the point. These factors may appear in different news messages. But if in several messages the same factors are revealed, then these messages should have the same topic. In particular, the topics of messages from example are *Japanese foreign commerce* and *activity of American administration*.

TABLE I.

Message 1
U.S. ASKS **JAPAN** TO END AGRICULTURE IMPORT CONTROLS **TOKYO**, March 3
The U.S. Wants **Japan**[1] to eliminate import controls on agricultural products within three years, visiting U.S. Under-Secretary of State for **Economic**[1] Affairs Allen Wallis **told**[2] Eishiro Saito, Chairman of the Federation of **Economic**[1] Organisations (Keidanren), a spokesman for Keidanren said. The spokesman quoted Wallis as saying drastic measures would be needed to stave off protectionist legislation by **Congress**[3] .Wallis, who is attending a sub-cabinet-level bilateral **trade**[1] meeting, made the remark yesterday in talks with Saito. Wallis was quoted as saying the **Reagan**[3] **Administration**[3] wants **Japanese**[1] cooperation so the **White House**[3] can ensure any U.S. **Trade bill**[1] is a moderate one, rather than containing retaliatory measures or antagonizing any particular country. He was also quoted as saying the U.S. Would be pleased were **Japan**[1] to halve restrictions on agricultural imports within five years if the country cannot cope with abolition within three, the spokesman said. **Japan**[1] currently restricts imports of 22 agricultural products. A ban on rice imports triggered recent U.S. Complaints about **Japan's**[1] agricultural policy.
End messge 1_

TABLE II.

Message 2
U.S. COMMERCE SECRETARY QUESTIONS FUJITSU DEAL **WASHINGTON**, March 3
Commerce Secretary Malcolm Baldrige said he felt a proposed takeover by **Japan's**[1] <Fujitsu Ltd> of U.S.-based Fairchild Semiconductor Corp, a subsidiary of Schlumberger Ltd <SLB>, should be carefully reviewed. He **told**[2] the Semiconductor Industry Association the deal would soon be discussed by representatives of several different **government**[3] departments. The **Reagan administration**[3] has previously expressed concern that the proposed takeover would make Fujitsu a powerful part of the U.S. **market**[1] for so-called supercomputers at time when **Japan**[1] has not bought any a American-made supercomputers. In addition, U.S. defense **officials**[3] have said they were worried semiconductor technology could be transferred out of the United States, eventually giving **Japanese**[1]-made products an edge in American high-technology markets for defense and other goods. Treasury Secretary James Baker recently **told**[2] a **Senate**[3] committee the proposed takeover would be reviewed by the cabinet-level **Economic**[1] Policy Council.
End messge 2

Evidently, factors reflect mutual meaning of the messages quite right. Here terms marked 1 are contained in the first factor, terms marked 2 are common words - contained in both factors and terms marked 3 are words contained in the second factor. One can see that this factor analysis is really nonlinear (it produces overlapping cluster) as there is nonempty set of common words.

4.2 Analysis of parliament voting

For the following analysis we used as data source results of roll-call votes in the Russian parliament in 2004 (INDEM Statistica, 2004). Each vote is represented by a binary vector with component 1 if the correspondent deputy voted affirmatively and 0 negatively. The number of voting during the year was 3150. The number of deputies (consequently the dimensionality of the signal space and the network size) was 430 (20 deputies who voted less than 10 times were excluded from the analysis).
By Boolean Factor analysis we found 5 factors in order according their highest Lyapunov function. First factor consists of 50 deputies and completely coincides with the fraction of the Communist Party (CPRF). Second factor consists of 36 deputies. All of them belong to the fraction of the Liberal-Democratic Party (LDPR) that has 37 chairs in the parliament in total. Thus one of the members of this fraction fell out of the corresponding factor. The third factor consists of 37 deputies. All of them belong to the fraction “Motherland” (ML) which consists of totally 41 deputies. Thus 4 of its members fell out of the factor. Next factor, compared with the list of deputies has shown that they correspond to the members of the fraction “United Russia” (UR). This fraction is the largest one, consist totally of 285 deputies, but it is less homogeneous. Therefore the Lyapunov function was relatively low.
The two remaining factors contain members of UR and independent deputies (ID). Factor with the higher value of Lyapunov function contains only members of UR and lower one – mainly ID but also members of UR. The general relation between the parliament fractions and obtained factors is shown in Table III.

TABLE III.
Relation between parliament fractions and factors

	1	2	3	4	5
CPRF	0 / 0	51/49	0 / 0	0 / 2	0 / 0
LDPR	1 / 2	0 / 0	36/ 35	0 / 0	0 / 0
ML	3 / 3	0 / 0	0 / 0	37/ 38	1 / 0
ID	1 / 14	0 / 0	0 / 1	0 / 1	15 / 0

The fit between the fractions and the factors was evaluated by F-measure (Riesbergen, 1979). Averaged over all fractions it amounted to 0.98.

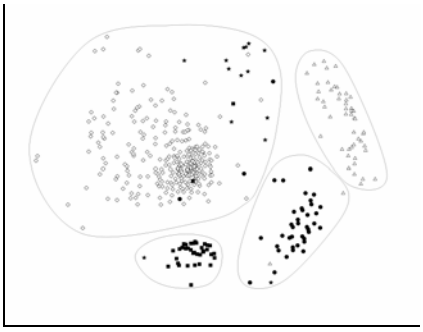

Fig. 1 *Two-dimensional map of voting parliament members. Thin lines - borders of clusters.* ♦ *- UR,* ▵ *– CPRF,* ■ *– LDPR,* ● *- ML,* ★ *– ID.*

We compared our results with those obtained using some traditional clustering methods (Kaufman and Rousseeuw, 2005; The Public Whip, 2006). First, we clustered the parliament members with the direct use of a similarity matrix. Similarity between two deputies was calculated by comparison of vectors of their voting. We used different measures of similarity: Euclidean distance, cosine, Jaccard and Dice. Both hierarchial and *k*-means clustering gave clusters far from parliament fractions: all fractions intersected in clusters and fraction LDPR could not be separated from ER at all.

Second, we performed mapping of parliament members by the method of multidimensional scaling. The results are shown in Fig. 1. This map was clustered. The borders of clusters are shown by thin lines. Generally, as factors obtained before, clusters coincide with parliament fractions except for independent deputies. The results of clustering and factorization are compared in the Table III.. The mean F-measure amounted to 0.95 that is slightly smaller than that obtained for factors.

5. Conclusions

The experimental results of the analysis of different data sets by the neural network method show that this method provides reasonable results compared with traditional attempts or their combination. The advantage of the neural method is its' less complicated usage and better scalability. We propose that they outperform traditional methods when applied to large data sets with high dimensionality.

References

Frolov A. A., Húsek D., Muraviev I. P. (1997) Informational capacity and recall quality in sparsely encoded Hopfield-like neural network: Analytical approaches and computer simulation. *Neural Networks*, 10**,** 845–855.

Frolov A. A., Husek D., Muravjev I. P. (2003) Informational efficiency of sparsely encoded Hopfield-like autoassociative memory. *Optical Memory and Neural Networks (Information Optics)*, 177–198.

Frolov A.A., Sirota M., Husek D., Muraviev I.P., and Polyakov P.Y. (2004) "Binary Factorization in Hopfield-like Neural Networks: Single-Step Approximation and Computer Simulations", Neural Networks World, Vol. 14, , pp. 139 - 152.

Goles-Chacc E., Fogelman-Soulie F. (1985) Decreasing energy functions as a tool for studying threshold networks. *Discrete Mathematics*, 261–277.

Gordon A. D.(1999). *Classification, 2nd Edition*. Chapman & Hall/CRC, Boca Raton..

Hopfield J. J. (1982) Neural network and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79, 2544–2548.

Husek D., Frolov A.A., Rezankova H., Snasel V., and Polyakov, P.Y. (2005) "Neural Network Nonlinear Factor Analysis of High Dimensional Binary Signals". In: *Proceedings of conference SITIS 2005*, University of Bourgogne, Dijon, pp. 86 - 89.

INDEM Statistica. (2004) http://www.indem.ru/indemstat Retrieved 7.7.2006.

Kaufman L. and Rousseeuw P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. (Wiley Series in Probability and Statistics). Wiley, New Jersey.

Newman D. J., Hettich S., Blake C. L., Merz C. J. (1998). UCI Repository of machine learning databases. Irvine, CA: University of California, [http://www.ics.uci.edu/~mlearn/MLRepository. html]

Perez-Vicente C. J., Amit D. (1989) Optimized network for sparsely encoded patterns, *J. of Physics A: Math. Gen.*, 22, 559 – 569.

Polyakov P., Frolov A. A., Húsek D. (2006) Hopfield Neural Network Based Binary Factor Analysis of Textual Data In: *Proceedings Of VIIIi All-Russian Scientific Conference "Neuroinformatics-2006"* (Ed.: Tumencev J.V.) - Moscow, MIFI Held: Moscow, January 24-27, 2006

The Public Whip. (2006) http://www.publicwhip.org.uk. Retrieved 7.7.2006.

Reuters. (2004) http://www.daviddlewis.com/resources/testcollections/reuters21578/. Retrieved 7.7.2006.

Riesbergen C. J. (1979) Information Retrieval, London: Butterworth, available also from http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html. Retrieved 7.7.2006.

Acknowledgement: The work was partly supported by the Institutional Research Plan AVOZ10300504 "Computer Science for the Information Society: Models, Algorithms, Appplications", by grant 201/05/0079 awarded by the Grant Agency of the Czech Republic and by grant No. IET100300414 granted by GA AS CR.