Combining different Clustering Techniques for Improved Knowledge Discovery

Olivera Grljević, Zita Bošnjak
Faculty of Economics, Subotica
University of Novi Sad
Segedinski put 9-11, 24000 Subotica, Serbia
{oliverag, bzita}@eccf.su.ac.yu

Abstract. *Application of different clustering techniques can result in different basic data set partitions emphasizing diversified aspects of resulting clusters. Since analysts have a great responsibility for the successful interpretation of the results obtained through some of the available tools, and for giving meaning to what forms a qualitative set of clusters, additional information attained from different tools is of a great use to them. In this article we presented the clustering results of small and medium sized enterprises' (SMEs) data, obtained in DataEngine, iData Analyzer and Weka tools for intelligent analysis.*

Keywords. Data mining, clustering, DataEngine, iData Analyzer, Weka

1 Introduction

The idea of Knowledge Discovery in Databases (KDD) is to search for relations and global schemes that exist in large databases and are hidden in the vast amount of data. Data mining, as the part of KDD, is the process of using one or more computational techniques in automated search for hidden information and relationships among data. As such, it represents infallible part of qualitative research. Knowledge discovered, through different data mining methods and techniques reveal behavioral patterns, profiles of entities, and similar regularities in data. Using solely statistical methods, qualitative data model can not be built. Besides large databases, sophisticated algorithms are needed, which are subject of knowledge discovery in databases.

As proven by now, each clustering algorithm, sometimes even the same algorithm applied several times on the initial dataset, can result in different basic dataset partitions, putting an accent on a specific aspect of the resulting clusters. Apart from diverse outputs, clustering algorithms use different visualization techniques to represent the derived clusters, which enable better insight into their structure and grouping relationships of similar entities. Furthermore, they can denote cluster centers, typical and least typical representatives of clusters, etc.

The selection of a subset of attributes in the database for clustering data, as well as the determination of the most adequate number of clusters, is under subjective appraisal of analysts. Furthermore, they have a great responsibility to carry out the interpretation of the results gained through some of the available tools successfully, and to give meaning to what forms a qualitative set of clusters. Consequently, additional information attained from different tools that support clustering techniques is of great use in clusters shaping.

Collecting and compounding various information about defined clusters, contributes to qualitative decision making on optimal cluster number and elements that constitute them. Consequently, to obtain as qualitative results as possible, and to facilitate cluster interpretation, analysts should combine different tools in the process of data clustering.

In our paper, we described a composite approach that implies diversity of tools and obtained results that significantly simplify the work of analysts in knowledge discovery, helps the interpretation of

This paper is the result of a research on the project titled: "Comparative Advantages of Intelligent Data Analysis Methods in Strengthening the Sector of Small and Medium Sized Enterprises", No. 114-451-01092/2008-01, funded by the Ministry of Sciences and Technology Development of Province of Vojvodina, Republic of Serbia.

results, and facilitates the derivation of detail and clear conclusions. We present the results of clustering small and medium sized enterprises' (SMEs) data in Vojvodina province using DataEngine, iData Analyzer and Weka tools for intelligent analysis. Each tool supports a different clustering algorithm.

2 Data mining tools overview

The goal of our research is to determine discriminators between successful and less successful enterprises, and to distinguish the profile of businesses that will succeed in their goals from those that are likely to fail. These tasks are classification and clustering tasks, respectively. Ref. [2] provides more detailed presentation of these problems. In this article we presented only the results of clustering techniques utilization, since they are common to all three tools we used, and are in compliance with the goals of our research.

Ref. [16] states that clustering is a process of grouping feature space vectors into classes in the self-organized mode. Cluster is a group of points in a multi-dimensional space. The points aggregated in such a way are closer to each other and to their "cluster center" than they are to the centers of other groups.

DataEngine (DE) software tool for intelligent data analysis is a very powerful tool that facilitates knowledge discovery in data. It combines statistical methods with neural networks technology, both supervised and unsupervised learning models, and fuzzy technology. Intelligent technologies DE supports are well proven in business, technology and academy work. In DE all data processing steps can be automated by graphical macro language and all models developed in DE can be incorporated into user's own programs (if they are built as Dynamic Link Libraries, for instance).

DE uses the Fuzzy C-Means (FCM) algorithm for partitioning a collection of points into a number of clusters. These data points are represented as feature vectors and are describing objects. The objects within a cluster show a certain degree of closeness or similarity. Objects are assigned to each cluster with a corresponding membership degree. The algorithm is using validity criteria to determine number of clusters in the data.

The FCM has several drawbacks that influence its performance. "The main drawback is from the restriction that the sum of membership values of a data point x_i in all the clusters must be one, and this tends to give high membership values for the outlier points." (Ref. [1]). Therefore, the algorithm has a problem in handling outlier points. The second limitation refers to the fact that membership of a data point in one cluster is directly related to its membership values in other cluster centers. Sometimes this leads to unrealistic results. This algorithm produces partial memberships in all the clusters for each point, which leads us to the third limitation, as partial membership of all data members moves clusters centers towards the center of all data points. The last limitations refers to FCM inability to calculate the membership value if the distance of a data point is zero.

In Ref. [18] it is stated that "The iData Analyzer (iDA) provides support for business or technical analyst by offering a visual learning environment, an integrated tool set, and data mining process support". iDA consists of a preprocessor for improving the quality of data, three data mining tools: unsupervised clustering, supervised learning and neural networks, and a report generator. iDA is an Excel add-on, so the user interface is Microsoft Excel. It uses first three rows of a spreadsheet to store the information about individual attributes. In this way, it states if the attribute has categorical or numerical value, if it should be used as input in model building or as an output attribute. There is also a possibility to declare certain attributes as unused or display-only, when they would not be used for building a model. Each column in MS Excel spreadsheet can represent an individual attribute.

The essential limitation of commercial version of iDA is that it can work with a single MS Excel spreadsheet, which allows maximum of 65536 rows and 256 columns. The version of iDA, which we have used, has even greater limitation regarding the dataset size – no more than 7000 data instances can be mined with this tool. The maximum size of an attribute name or value stored in one cell is 250 characters. The last limitation is that RuleMaker in iDA will not generate rules if the number of derived classes exceeds 20.

An exemplar-based data mining tool (ESX), which builds a concept hierarchy to generalize data can, as stated in Ref. [18], "help create target data, find irregularities in data, perform data mining, and offer insight into the practical value of discovered knowledge". ESX will not make statistical assumptions about the nature of mined data. Furthermore, it can emphasize certain inconsistencies and unusual values in dataset. If ESX is performing supervised classification, it can provide information about those instances and attributes which could classify in the best fashion new instances of unknown origin. When performing unsupervised clustering, ESX incorporates a globally optimizing evaluation function that encourages a best instance clustering. In contrary to DataEngine, iDA can work both with categorical and numerical data values.

Waikato Environment for Knowledge Analysis - Weka is suite of Java class libraries and it implements many acknowledged machine learning and data mining algorithms. In contrary to DE and iDA, algorithms in Weka can be applied either directly to a dataset or can be called from Java code. It contains tools for preprocessing, classification, regression, clustering, association rules and visualization. It is also suited for developing new machine learning schemas. Pros for using Weka tool are the following: it covers the entire machine learning process, it facilitates comparison of the results of different algorithms implemented, it accepts one of the most widely used data formats as input – ARFF format, there are flexible APIs for programmers, and customization possibilities. Weka has also some deficiencies: it requires Java Virtual Machine to be installed for its execution, and visualization of mining results is not possible.

Weka tool implements clustering methods as k-Means, EM, Cobweb, X-means, FarthestFirst, and others. We decided to use simple k-Means algorithm as it is one of the oldest and most widely used clustering algorithms.

K-Means algorithm is a prototype-based, partitioning technique that attempts to find a user-specified number of clusters (k), which are represented by so called centroids. Centroid is usually the mean of a group of points and is typically applied to objects in a continuous n-dimensional space [Ref. 20]. It is a very simple and fast algorithm. Since k-means requires that the user knows the exact number of clusters (k) in advance, and usually this number is not obvious, determining the initial value of k is a major difficulty in using this algorithm. Furthermore, a lack of explanation requires additional analysis by a supervised learning model.

3 Data understanding

The goal of our research was to discover knowledge hidden in small and medium sized enterprises' (SMEs) data, by means of intelligent data analysis and in that way to support the development of this sector. The SMEs data were provided by four Regional Agencies for the Development of Small and Medium Sized Enterprises and Entrepreneurship from province of Vojvodina. The data was collected in 2006. by means of the questionnaire these Agencies provided. The questions in the questionnaire were divided into two groups. The first group aimed to collect general enterprise data. The second group of data was formed by answers of individual enterprises to the questions related to business itself, technical, technological and financial aspects, market conditions and distribution, administrative and legislative conditions, human resources, business connectivity, and the need for non-financial services.

The final data collection consists of 2365 records on SMEs in the province of Vojvodina. Each data record is described with more than one hundred attributes. The data was originally stored in MS Access format and contained many missing data. Therefore, there was a need for qualitative data transformation into a format required by each data analysis tool we used in our research. Also, in the data preprocessing phase, many of initial attributes were removed from further analysis (data preprocessing is described in more detail in [11]). The resulting set of data was divided into subsets, and different tools, data mining methods and techniques were used for their analysis. In this paper we presented the data analysis results, using different clustering techniques. At this point, it is essential to emphasize the fact that the quality of collected data was poor and that we faced many challenges during the data mining. Consequently, there are some limitations in applicability of revealed knowledge (these challenges and limitations are described in more detail in [2]).

4 Data analysis

The first analysis we carried out covers the data about problems enterprises cope with, innovations they have conducted in the previous two years, over aging of fixed assets, percentage of capacity utilization, and the ownership structure. Firstly we developed the clustering model which divides the SMEs according to the main problems they were facing in everyday business operations (lack of available funds, complex administrative and legislative regulations, disharmony with standards, insufficient market information, insufficient information on technologies, unavailability of qualified work force, and human resources development). In DE tool, a possibility of cluster analysis is available, where we used a partition coefficient as a validity measure to determine the best number of clusters. For the same purpose, another two validity criteria can be used, a proportion exponent and a classification entropy. These are three known criteria by which fuzzy clustering can be judged. Also, their values can be presented in a form of a graph. We also inspected these criteria by putting in relationship the partition coefficient (denoted pc) and the classification entropy (pe), shown in Fig. 1. As stated in Ref. [10] both of these validity criterions tend towards monotone behavior depending on the number of clusters. Therefore, to determine the optimal number of clusters (c) we had to look for the number of clusters at which these values have a kink, a so called "elbow criterion". The graph presented on Fig. 1 shows that, according to these two criteria, the best partitioning of the SMEs data regarding the problems they faced in everyday business activities is achieved with four clusters. The same number of clusters was obtained using cluster analysis in DE. Fig. 2 represents these SMEs groups in DE.

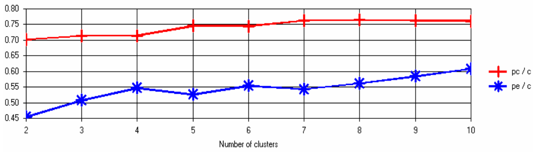


Figure 1. Optimal cluster number determination in DE tool

Analyzing a resemblance score, as a main indicator of successfulness of clustering process, and of the goodness of the model developed in iDA tool, we witnessed that the iDA clustering technique partitioned the same data into the same optimal number of clusters. Furthermore, we used this knowledge to set the initial number of clusters in Weka. Although each tool uses different clustering algorithm and therefore the structure of obtained clusters is also different, we were able to gain additional information analyzing defined clusters in each tool. This information was very valuable for better understanding of sector SMEs structure, regarding the problems they were facing with, and for determining other relationships.

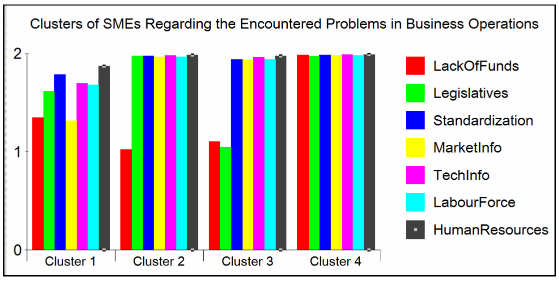


Figure 2. DataEngine clustering model regarding the main problems in business operations

We were able to produce parallel and visualized presentation of the distribution of SMEs answers on questions about each problem only in Weka tool. This way we could have clear inspection of SMEs question structure, shown in Fig. 3. This figure clearly shows that the majority of enterprises stated that they did not experience any problem in their business operations. The fact that clustering results of DE and iDA provide the same structure of one cluster that is comprised of SMEs that had not recognized any of the above listed threats as a serious one to their business operations (rank 2 in Fig. 2) is not surprising. This cluster in iDA accounts 84% of all instances from the initial dataset. iDA clustering resulted in another cluster, that comprises solely of those enterprises that stated all treats as a problem in their business activities. Despite the small number of such enterprises (just eight of them), this cluster should be considered carefully, as there is a possibility that lot of similar enterprises have not been able to cope with all difficulties they encountered, and therefore have not manage to survive in business. On Fig. 3 it can be seen that the most commonly occurring problems are lack of available funds and complex administrative and legislative regulations. Also, it can be observed that lack of available funds is equally distributed among enterprises. On Fig. 2 it can be seen that enterprises belonging to other clusters have also recognized the lack of funds as a serious threat (rank 1). Additional information gained in iDA, that was not available in other two tools, refers to the most commonly occurring attributes in defined clusters, and the typical representatives of each cluster. iDA provides the list of all instances belonging to one cluster ranked from the most typical for that cluster to the least typical with associated typicality score. Fig. 4 shows the most commonly occurring attributes in formed clusters in iDA. iDA revealed interesting information about typical representatives of clusters number two and three. The most typical representatives of cluster 2 are those enterprises that have experienced problems with insufficient market information and insufficient information on new technologies. An interesting finding is that these enterprises have old generation fixed assets, and the percentage of capacity utilization average is from 70 to 80%. The most typical representatives of the third cluster are those enterprises that stated their main problems as lack of available funds and insufficient market information. Their fixed assets vary from a new generation to out dated ones. Enterprises that have problems also with regulation, information on new technologies, available fork force, and human resources, belong to this cluster, as well.

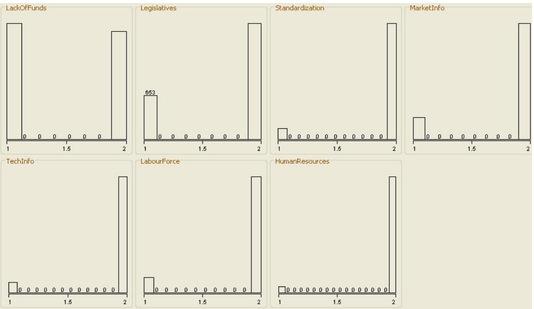
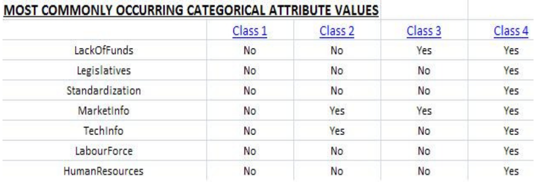


Figure 3. Distribution of answers on question about main problems in Weka

Surprising was the finding that membership of SMEs in domestic/foreign business associations, as well as their involvement in industry clusters[2] had no influence on overcoming the problem of insufficient funds, despite the fact that both business associations and industry clusters are established primarily for this reason.



MOST COMMONLY OCCURRING CATEGORICAL ATTRIBUTE VALUES

	Class 1	Class 2	Class 3	Class 4
LackOfFunds	No	No	Yes	Yes
Legislatives	No	No	No	Yes
Standardization	No	No	No	Yes
MarketInfo	No	Yes	Yes	Yes
TechInfo	No	Yes	No	Yes
LabourForce	No	No	No	Yes
HumanResources	No	No	No	Yes

Figure 4. Most commonly occurring categorical attribute values in iDA[3]

[2] The distinction between the term “cluster” in a sense of data mining output and the term “cluster” in a sense of grouping of similar business entities, that share a common business goal, should be made. Therefore, in the paper the latter was replaced by term “industry cluster”.

[3] In iDA tool, terms “class” and “cluster“are used as synonyms.

The predominant ownership structure among analyzed SMEs is the private one. Those enterprises that have private ownership structure that is equal or greater than 10000 and less than 100000 dinars, innovated in the previous two years their organization and technology. Those SMEs that have private ownership structure greater than zero and less than 10000 dinars innovated their products or services. According to our findings, SMEs that had certain innovation activities, stated also insufficient stimulating financial sources or lack of such, as their main problem. Fig. 5 represents cluster structure regarding innovations, in DE tool. The similar structure was obtained in iDA and in Weka tools.

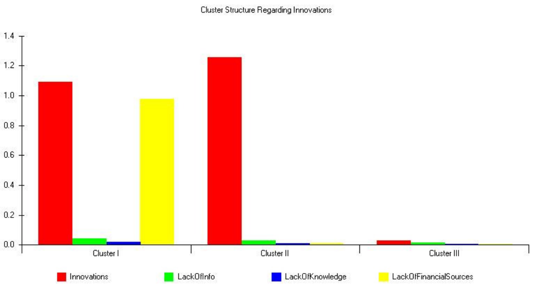


Figure 5. Clusters structure regarding innovations in DE tool

Given the fact that Weka has excellent visualization possibilities, we used this advantage of Weka tool to display frequency of occurrence of answers within clusters, as shown in Fig. 6. Such presentation of clusters structure was not possible in other two tools. This figure shows that only few enterprises stated other problems than already mentioned, as a reason for not being engaged in innovation activities. The red spots indicate that SMEs has recognized certain problem, blue spots indicate otherwise.

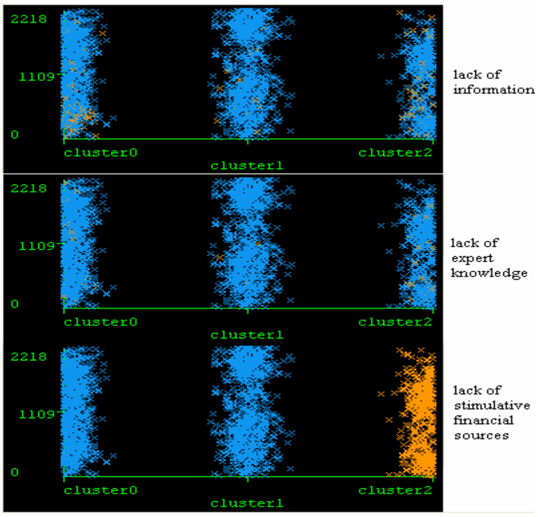


Figure 6. Frequency of occurrence of each answer in derived clusters in Weka

5 Conclusion

In this article, we described the application of different clustering techniques for analyzing data on small and medium sized enterprises, in order to obtain results which could support the development of SMEs sector.

However, during data analysis we aspired also to develop a classification model which could classify SMEs based on seven input attributes related to difficulties in everyday business operations, into one of the three predefined classes:

1 – SMEs with outdated equipment
2 – SMEs with moderately outdated equipment
3 – SMEs with “new generation” equipment.

Unfortunately, in DE tool we were unable to build such a model using the Multilayer Perceptron classification model. We were able to develop such a model in Weka tool, but the classification error, i.e. the percentage of incorrectly clustered instances was too high. The main reason for that was a poor quality of collected data. Namely, SMEs representatives were not motivated sufficiently to provide all the required answers by the questionnaire.

Unsuccessful data analysis attempts, such as the one mentioned above, lead us to the conclusion that the composite approach to data analysis process, that implies diversity of tools, could not help in achieving each data mining goal. Despite this fact, we managed to take advantage of the utilization of three tools – DE, iDA, and Weka, when clustering tasks are in question. Application of each tool added additional information to the previously discovered knowledge. We presented these results in short in this paper.

Agencies for the Development of Small and Medium Sized Enterprises and Entrepreneurship could use the relationships discovered in SMEs everyday business data, regarding the problems they are coping with, to devise programs to better suite the needs of SMEs, and to enable, in that way, the further development of SMEs sector. The discovered knowledge could also be applied for decreasing a trend of closing enterprises, by early recognition of those SMEs that are similar to SMEs recorded to be unsuccessful in the previous few years, and upon recognition, helping their businesses to survive and improve.

References

[1] Binu T., Raju G., Sonam W.: **A Modified Fuzzy C-Means Algorithm for Natural Data Exploration**, Proceedings of world academy of science, engineering and technology, Volume 37, January 2009, Waset.org, 2009 , pp. 478-481.

[2] Bošnjak Z., Grljević O., Bošnjak S.: **CRISP-DM as a Framework for Discovering Knowledge in Small and Medium Sized Enterprises**, unpublished.

[3] Bošnjak Z., Grljević O.: **Data Mining as a Mean for Devising Actions for Development of the Sector of Small and Medium Sized Enterprises**, unpublished.

[4] Bošnjak Z., Bošnjak S., Stojković M.: **Application of Fuzzy Clustering for Searching Trends in Data – The Public Transport Company in Subotica Case Study**, Proceedings of EUROFUSE 2005, 15th -18th Jun, Belgrade, **Serbia, 2005.**

[5] Bratko I.,. Kubat M, R.S. Michalski: **Machine Learning and Data Mining: Methods and Application**, John Wiley & Sons Inc., 1998.

[6] Cahlink G.: **Data Mining Taps and Trends**, Government Executive Magazine, http://www.povexec.com/tech/articles/1000managetech.html, October 1, 2000.

[7] Cios K.J., Kurgan L.A., Swiniarski R.W., Pedrycz W.: **Data Mining: A Knowledge Discovery Approach**, Springer Science+ Business Media LLC, 2007.

[8] **DataEngine**, available at http://www.mitgmbh.de

[9] **DataEngine – User Guide**, MIT, Germany, 2008.

[10] **DataEngine Tutorials and Theory**, MIT GmbH, Aachen, Germany, 1997.

[11] Grljević O., Bošnjak Z.: **Primena CRISP-DM Metodologije u Analizi Podataka o Malim i Srednjim Preduzećima (CRISP-DM Methodology Utilization in Preprocessing Small and Medium Sized Enterprises Data)**, Book of proceedings of XXXV symposium on OR, SYM-OP-IS, Septembar, Soko Banja, Serbia, 2008, pp. 275-279.

[12] Harrison P.G., Llado C.M.: **Performance Evaluation of a Distributed Enterprise Data Mining System Source**, Lecture Notes In Computer Science, Vol. 1786, Springer-Verlag, London, 2000, pp. 117-131.

[13] **iData Analyzer**, available at http://www.infoacumen.com

[14] Jiawei H., Kamber M.: **Data Mining Concepts and Techniques**, Morgan Kaufman Publishers, San Francisco, 2001.

[15] Komem J., Schneider M.: **DataEngine Tools for Inteligent Data Analysis and Control**, Data Mining and Knowledge Discovery Handbook, Springer Science+Business Media Inc., 2005, pp. 1371-1377.

[16] Liao T. W., Triantaphyllou E.: **Recent Advances in Data Mining of Enterprises Data: Algorithms and Applications**, Vol. 6, Series on Computer and Operations Research, 2008.

[17] Nemati H.R., Barko C.D.: **Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance**, Idea Group Inc (IGI), 2003.

[18] Roiger R. J., Geatz M. W.: **Data Mining: A Tutorial – Based Primer**, Addison Wesley, USA, 2003.

[19] Seifert J.W.: **Data Mining: An Overview**, CRS Report for Congress, http://www.fas.org/irp/crs/RL31798.pdf, December 16, 2004.

[20] Tan P-N. Steinbach M.,Kumar V.: **Introduction to Data Mining,** Addison Wesley, USA, 2007.

[21] **Waikato Environment for Knowledge Analysis**, available at http://www.cs.waikato.ac.nz/ml/weka/

[22] Witten I.H., Frank E.: **Data Mining: Practical Machine Learning Tools and Techniques**, Elsevier Inc., 2005.