Different Clustering Techniques – Means for Improved Knowledge Discovery

Olivera Grljević[1] - Saša Bošnjak[2] - Zita Bošnjak[3]

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 indivisible 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.

[1] Olivera Grljević, MSc, University of Novi Sad, Faculty of Economics Subotica, Serbia
[2] Saša Bošnjak, PhD, associate professor, University of Novi Sad, Faculty of Economics Subotica, Serbia
[3] Zita Bošnjak, PhD, full professor, University of Novi Sad, Faculty of Economics Subotica, Serbia

Acknowledgement: 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.

Besides large databases, sophisticated algorithms are needed, which are subject of knowledge discovery in databases.

As proven by now (Liao- Triantaphyllou 2008), (Harrison-Llado 2000) 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 of 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 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. Techniques used in the empirical examination

The broader goal of our research was 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. (Bošnjak et al. 2009) 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.

In Komem- Schneider (2005), Bratko et al. (1998), Jiawei-Kamber (2001) it is defined 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. Within our research, we have created data models by c-means algorithm, improved fuzzy c-means algorithm and Kohonen neural networks for clustering tasks.

C-Means algorithm is a prototype-based, partitioning technique that attempts to find a user-specified number of clusters (c), which are represented by centroids. Centroid is usually the mean of a group of points and is typically applied to objects in a continuous n-dimensional space, (Tan et al. 2006), (Witten-Frank 2005). It is a very simple and fast algorithm. Since c-means requires that the user knows the exact number of clusters in advance, and usually this number is not obvious, so determining the initial value of c is a major difficulty in using this algorithm. Furthermore, a lack of explanation requires additional analysis by a supervised learning model. In a crisp C-means algorithm, each entity belongs to only one cluster, not being the case in numerous real world situations, and hence the algorithm is facing a limited usability.

One improvement of the C-means clustering algorithm incorporates the theory of fuzzy sets, resolving the single-membership problem by measuring the degree of membership of all entities to each cluster, by the membership function. However, the fuzzy C-means (FCM) algorithm still has several drawbacks that influence its performance. (Binu et al. 2009) it is stated that "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." The second limitation refers to the fact that membership of a data point in one cluster is directly related to its membership in other clusters, and also the partial membership of all data members moves clusters' centers towards the center of all data points, producing sometimes unrealistic results. Consequently, additional information attained from different data analysis tools that support clustering techniques is of great use in clusters shaping.

3. Data mining tools overview

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.

(Roiger- Geatz 2003) 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 Roiger- Geatz (2003), "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 with both 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 C-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. We decided to use the simple c-means clustering algorithm, as it is available in all three selected data mining tools and is generally in wide use.

4. 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.

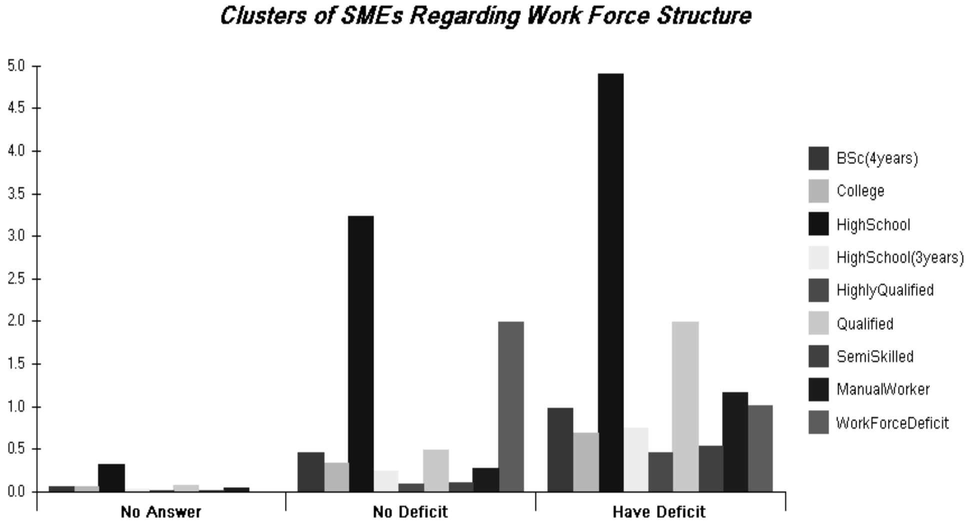
The final data collection consists of 2365 records on SMEs in the province of Vojvodina. Each data record is described with more than one hundred attributes. The data was originally stored in MS Access format and contained many missing data. Therefore, there was a need for qualitative data transformation 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 (Grljević- Bošnjak 2008).

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 (Bošnjak et al. 2009).

5. Data analysis

The analyses we have carried out, and the results presented in this article, refer to human resources data. They consist of data concerning an employee's qualification structure (B.Sc level, College, High School – 4 years, High School - 3years, Highly Qualified, Qualified, Semi-qualified, Without qualification) and the deficit of adequate work force, experienced by SMEs (facing/not facing such a deficit). Firstly we developed a clustering model which divides SMEs according to the structure of work force. DE tool offers a possibility of cluster analysis, where we used the partitioning coefficient as a validity measure to determine the best number of clusters. For the same purpose, other two validity criteria can be used: proportion exponent and classification entropy. These are three known criterions by which fuzzy clustering can be judged. Also, they can be presented in a form of graph. We inspected these criteria, putting in relationship the partitioning coefficient (pc) and the classification entropy (ce). As stated in (MIT GmbH, 1997) both of these validity criteria 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, so called "elbow criterion" (the best number of clusters according to this criteria was 3). DE offers an additional functionality, called cluster analysis that automatically determines the optimal cluster number. According to cluster analysis, the optimal number of clusters was also 3. Combining these two possibilities, we found that the best partitioning of the SMEs data regarding work force structure and work force deficit an enterprise is facing with, is achieved with three clusters. Fig. 1 represents these SMEs groups in DE.

Figure 1. DataEngine clustering model regarding the work force deficit



Source: Own creation

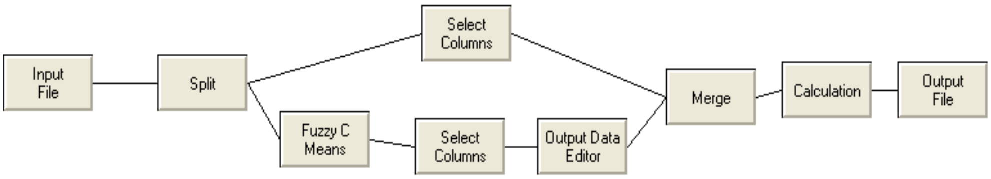
Analyzing the resemblance score, as a main indicator of successfulness of clustering process in iDA tool, and the goodness of the model developed, we revealed that usage of clustering technique in iDA to partition the same data, resulted in the same optimal cluster number. Furthermore, we used this knowledge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

Out of those enterprises that cooperated and gave answers to the questions, majority stated that they do not have a work force deficit. Also many SMEs did not provide an exact number of employees distributed amongst different levels of qualification. Given the previously stated fact, the generated cluster with all ranks ranging around zero (Fig. 1) is fully justified. Each tool we have used generated one such cluster and two others. One of them consists of SMEs facing work force deficit, while the other contains SMEs that declared they do not have such a deficit (rank 1 and 2 on Fig. 1, respectively).

Additional information gained in iDA, that were not available in other two tools, refer to 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 scores. The most typical representatives of cluster denoted NoDeficit are those enterprises that employ one employee with BSc level (four years study) of qualification, one employee with academic level of education and five or six employees with high school level of qualification. The most typical representatives of cluster HaveDeficit are those enterprises that employ one employee with BSc (four years study), and one with academic level of education, two to five employees with high school, and one highly qualified, and one qualified worker. We could conclude that, according to the most typical representatives of each cluster, a relationship between employee's level of education and deficit of qualified workforce is not relevant.

DE offers the possibility to automate data processing steps using graphical macro commands which are in a form of function blocks. Function blocks are placed on a card that allows their easy configuration and connection. Each card should contain at least one input and one output in the form of data or a graph. The processing steps are placed between input and output. We used this DE tool to inspect in more detail the structure of expressed deficit related to the level of qualifications. Fig. 2 shows a card that classifies all data according to the determined optimal cluster number, as we previously concluded it as three. After the correct classification of input data, this card will allow us to select several columns which will serve as input to macro block called calculation. This macro block provides the code that will select only those SMEs that surely have work force deficit, and data of such SMEs, together with their industry code, will be placed into in beforehand prepared data file.

Figure 2. Automated selection of SMEs with work force deficit



Source: Own creation

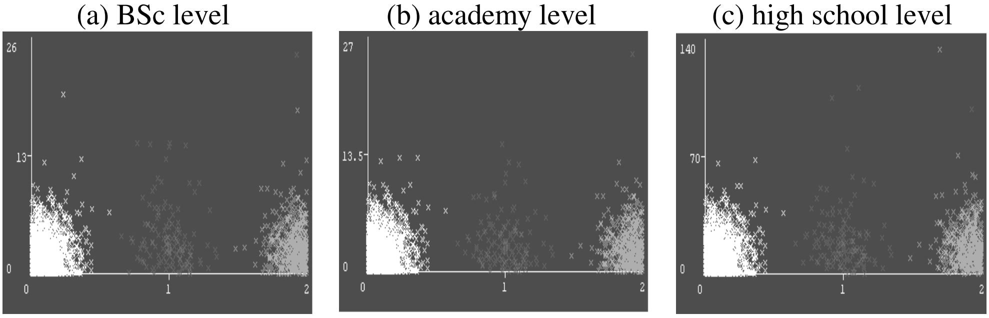
The goal was to inspect in more detail the work force deficit across different industry branches. The knowledge discovered revealed that 108 SMEs were facing work force deficit. Additional analysis of these data led us to the conclusion that SMEs come from very wide range of industry sectors, 60 different ones in total. The final conclusion we have driven was that the observed deficit was not sector specific.

After such a conclusion, we were peculiarly interested in relationships between work force deficit and different towns or municipalities in province of Vojvodina. Therefore, we did some further analysis. The presented card was modified in such a way that additional information was added to the already extracted and grouped ones. This information refers to the zip codes of towns where SMEs facing deficit located their businesses. In this way we gained better insight into desired relationships. The results show that majority of SMEs having work force deficit are from West-Bačka and South-Bačka administrative districts, and mainly from Novi Sad and Sombor municipalities. Therefore, we could conclude that, as we have suspected, the region and the deficit of work force were related.

As Weka offers some irreplaceable visualization functionalities, we used them to inspect further human resources aspect of SMEs. Figure 3 consists of three plots marked with (a), (b) and (c), displaying a frequency of occurrence of answers within clusters. Such presentation of cluster structure was not possible in other two tools. X axis represents a work force deficit, while Y axis represents the number of employees. This is displayed for 3 different levels of qualification: bachelor, academic and high school level of education as three highest levels. Other levels of qualification are omitted from this figure. The dots on graphs (a), (b) and (c) around (0,0) are those employees that answered with HaveDeficit, the dots in the middle of each graph represents those employees that answered with NoDeficit, and the cluster of dots on the right hand side represents employees that have not provided any answer. As can be seen in Fig. 3, SMEs that have fewer employees commonly face the deficit of adequate work force, since marks are concentrated around zero. The same trend can be observed when we visualize other levels of qualification with remarked deficit.

Generally speaking, the structure of work force in SMEs in province of Vojvodina, according to the level of qualification, is devastating. Probably one of the main reasons for this is the fact that SMEs are mainly family businesses that are inherited or developed irrespectively to the level of qualification of their owner(s) or employees. This is what led SMEs to a situation where most of directors are highly uneducated people (table 1).

Figure 3. Weka Cluster Visualization


(a) BSc level (b) academy level (c) high school level

Source: Own creation

Table 1. Level of SMEs directors' education

Director – education	
PhD	45
MSc	37
BSc (4 years)	346
Academy	237
High school (4 years)	77
High school (3 years)	809
Primary school	146
Other	17
No answer	628

Source: Own creation

Table 2 illustrates a number of employees according to their qualification structure. It shows that SMEs in province of Vojvodina mainly employ workers with high school level of education.

SMEs in province of Vojvodina are facing many different problems in everyday business, such as lack of available funds, complex administrative and legislative regulations, disharmony with standards, insufficient market information, insufficient information on technologies. Results of data analysis about problems SMEs cope with, innovations they have conducted in the previous two years, over - aging of fixed assets, percentage of capacity utilization, and ownership structure are presented in more details (Grljević- Bošnjak 2009).

Table 2. Employees qualification structure

Number of employees according to qualification structure	
PhD	679
Academy	526
High school	3962
High school (3 years)	550
Highly qualified	238
Qualified	1257
Semi-qualified	299
Unqualified	717

Source: Own creation

6. Conclusion

In this article, we described the application of different clustering techniques of small and medium sized enterprises, which could support the development of SMEs sector. The results presented refer to work force data, and unavailability of qualified work force, and human resources development.

During the research and analysis of data we concluded that a composite approach to data analysis process, that implies diversity of tools could not help in achieving each and every 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. Each tool has added additional information to the previously discovered knowledge. We presented these results in short in this paper, while additional results are available (Grljević- Bošnjak 2009).

The conducted analysis and the results presented in this paper are merely the starting point for further analysis. Additional analyses are needed to reveal, if possible, an exact reason of work force deficit. Agencies for the Development of Small and Medium Sized Enterprises and Entrepreneurship could use the results we described in this article, to create more successful employment policies and to maintain balance between supply and demand on work force market.

References

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

Bošnjak Z. – Grljević, O. – Bošnjak, S. 2009: *CRISP-DM as a Framework for Discovering Knowledge in Small and Medium Sized Enterprises*. Scientific Bulletin of „Politehnica" University of Timisoara, Romania, Transactions on Automatic Control and Computer Science, Vol 10. Fasc. /2010.

Bošnjak, Z.- Bošnjak, S.- Stojković, M. 2005: *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.

Bratko, I. - Kubat, M. – Michalski, R.S. 1998: *Machine Learning and Data Mining: Methods and Application*. John Wiley & Sons Inc, New York.

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

MIT GmbH, 1997: *DataEngine Tutorials and Theory*, Aachen, Germany.

MIT GmbH, 2008: *DataEngine – User Guide*, Aachen, Germany, available at http://www.dataengine.de/

Grljević, O. – Bošnjak, Z. 2008: *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, pp. 275-279.

Grljević, O. - Bošnjak, Z. 2009: *Combining different Clustering Techniques for Improved Knowledge Discovery.* Central European Conference on Information and Intelligent Systems, CECIIS 2009, Zagreb, Croatia.

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

iData Analyzer, available at http://www.infoacumen.com

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

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

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

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

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

Tan, P.-N. - Steinbach, M. – Kumar, V. 2006: *Introduction to data mining*, Addison Wesley, USA.

Waikato Environment for Knowledge Analysis. available at http://www.cs.waikato.ac.nz/ml/weka/

Witten, I.H. – Frank, E. 2005: *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier Inc., San Fransisco.