Perspectives of Innovations, Economics & Business, Volume 3, 2009
www.pieb.cz

INTELLIGENT DATA ANALYSIS - SUPPORT FOR DEVELOPMENT OF SMES SECTOR *

OLIVERA GRLJEVIC,
SASA BOSNJAK, PH.D.,
ZITA BOSNJAK, PH.D.

Faculty of Economics Subotica
University of Novi Sad, Serbia

JEL Classifications: C8, L2

Key words: Intelligent data analysis, CRISP-DM, clustering, small and medium enterprises.

Abstract: The paper studies possibilities of intelligent data analysis application for discovering knowledge hidden in small and medium-sized enterprises' (SMEs) data, on the territory of the province of Vojvodina. The knowledge revealed by intelligent analysis, and not accessible by any other means, could be the valuable starting point for working out of proactive and preventive actions for the development of the SMEs sector.

ISSN: 1804-0527 (online) 1804-0519 (print) PP. 57-58

Introduction

Methods and techniques of intelligent data analysis, also known as data mining algorithms, have become an important research area, because their joint application with traditional data analysis can reveal knowledge - hidden relations, behavioral patterns, entity profiles, and similar regularities in data stored in large databases or warehouses. In the paper we described our efforts to take advantage of the intelligent data analysis and discover some knowledge hidden in small and medium sized enterprises' (SMEs) data, with the aim to support the development of SMEs sector. The data collection consisted of 2365 records, each containing data on one SME in Vojvodina province. The data were collected by means of distributed questionnaires, provided by four Regional Agencies for the Development of Small and Medium Sized Enterprises and Entrepreneurship.

CRISP-DM as a framework for knowledge discovery

Knowledge discovery is a very intricate, uncertain and time consuming process. Therefore, of utmost importance is to follow the existing methodological guidance, among which the CRISP-DM (the CRoss-Industry Standard Process for Data Mining) is the most favored one. It comprises of the following tasks: (a) business understanding; (b) data understanding; (c) data preparation; (d) modeling; (e) evaluation; and (f) deployment. Generally, these tasks follow each other as subsequent phases, but within this main stream, many iterative cycles can be observed. In our research we also used this methodology.

Crucial for the success of intelligent data analysis was the second phase of CRISP-DM methodology, comprising of data formatting, description, exploration, and data quality verification, because the data source was partially erroneous and of poor quality. The source data were originally stored in MS Access format, but they were further transformed into appropriate input to intelligent data analysis tools. During the exploration of data on SMEs, some interesting things were observed, such as the fact that there were almost three times less female than male directors in SMEs in Vojvodina (24.06% vs. 64.27%). Such insights into the sector of SMEs had to be treated with precautions, due to the large number of missing data (for e.g. 12% of values for the director's gender were missing). In the phase of data cleansing, missing values were either replaced by some neutral value, or the records with missing data were excluded from further analysis. Visualization techniques were also very useful in "outliers" detection, and for verifying or rejecting initially stated hypothesis on data dependences. For e.g. the hypothesized correlation between equipment maturity and lack of investments into new technologies had to be rejected (r = -0.20295). This was in contrast with our expectation that SMEs that had outdated equipment invested less in the last 5 years than SMEs having contemporary equipment. In subsequent steps of the Knowledge Discovery in Databases (KDD) process, the unique set of data was divided into subsets and different data mining methods and techniques were used for their analysis, as described in the sequel.

The challenges and limitations of data modeling

The modeling phase of CRISP-DM methodology includes the application of different machine learning techniques, with wide scale of tunable parameters, each. The stress in our investigations was on devising discriminators among successful and less successful enterprises, to describe the general profile of businesses that were likely to fail in achieving their goals, to select the attributes of high predictability in forecasting future business gains/losses, etc. Within our research, we have created data models by fuzzy c-means algorithm and Kohonen neural networks for clustering

[*] 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.

tasks, and Multilayer Perceptron (MLP) neural networks for classification and forecasting tasks. As data mining is known to be a time-consuming, laborious endeavor, without guaranties that interesting and potentially useful patterns will be revealed, we were prepared to the creation of large number of data models and trial-and-error approach to data analysis. Some models, such as the clustering model which divided 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, etc.) resulted in interesting findings. By this model, enterprises were clustered into 4 groups. One of the clusters comprised of SMEs that had not recognized any of the above listed threats as a serious one to their business operations, while all other enterprises had recognized the lack of funds as a serious threat. Surprising was the finding that membership of SMEs in domestic/foreign business associations, as well as their involvement in industry clusters 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.

Some other models, like the MLP classification model, developed with the aim to classify SMEs based on seven input attributes related to difficulties in everyday business operations, into predefined classes: SMEs with outdated equipment, SMEs with moderately outdated equipment, and SMEs with “new generation” equipment, could not be built. After number of trials with different MLP configurations and various learning methods offered in applied tools, we concluded that the defined output variable was not dependant on the selected input attributes in any way, i.e. there was no hidden relation we were hopeful to find.

Some data models were impossible to build because of some manifestations of poor data quality that became obvious only during the modeling phase of CRISP-DM methodology. The clustering model we tried to build for investigating the similarities/dissimilarities between SMEs concerning legal and administrative limitations to the development of business operations was one of these.

Knowledge discovery improvement by utilization of diversified data analysis tools

Data analysts have a great responsibility to carry out the interpretation of the results obtained by data mining, and to give a meaningful explanation of observed relations. That is why we selected a composite approach to data analysis that implies an application of diverse tools, methods and techniques for data mining. Within the composite approach to data analysis process, we managed to take advantage of the utilization of DataEngine (DE), Intelligent Data Analyzer tool (iDA), and Waikato Environment for Knowledge Analysis (Weka) tool.

The combination of three different data mining tools proved to be superior to the application of one single data mining tool. For e.g., the clustering model, that partitions SMEs according to business problems they were coping with, was firstly developed in DE tool. We used a relation of the partitioning coefficient and the classification entropy as a validity measure, to determine the optimal number of clusters as 4. The iDA clustering technique partitioned the data into the same optimal number of clusters. We took advantage of this knowledge to set the initial number of clusters in Weka and conducted the clustering. The visualization of clustering results was different in all three tools, providing additional information on defined clusters to data analysts.

Conclusion

Although knowledge discovery in data, and data mining as its integral part, are indispensable for data analysts, it is true only in case of good quality input data. In the described case of the analysis of SMEs data, it was an unpleasant truth that the goal and importance of the knowledge discovery endeavor were not recognized or fully understood by SMEs representatives and that they were not motivated sufficiently to provide all the required data. Consequently, lots of source data were missing or erroneous. We managed to overcome this obstacle by utilization of multiply tools when clustering tasks were in question, and despite the limitations we found some very interesting and unexpected relations between particular subsets of attributes describing SMEs in Vojvodina. Agencies for the Development of Small and Medium Sized Enterprises and Entrepreneurship could use the discovered relationships in SMEs everyday business activities to structure programs to better suite the need of SMEs and to ensure in such a way further development of their businesses and to decrease the trend of closing such enterprises, which was evident in the previous few years. Also, the revealed knowledge could be used to create more successful employment policies and to maintain balance between supply and demand on the work force market.

References

Cahlink, G., 2000. “Data Mining Taps the Trends”, Government Executive Magazine, http://www.govexec.com/tech/articles/ 1000managete ch.html.

Cios, L., Kurgan, K., Swiniarski, R., Pedrycz, W., 2007. Data mining: a knowledge discovery approach, Springer Science + Business Media LLC, Heidelberg.

Grljević, O., Bošnjak, Z., 2008. “CRISP-DM Methodology utilization in preprocessing small and medium sized enterprises data”, XXXV Symposium on OR, SYM-OP-IS 2008, ISBN: 978-86-7395-248-2, pp.275-279, Belgrade, Serbia.

Harrison, P.G., Llado, C.M., 2000. “Performance evaluation of a distributed enterprise data mining system source”, Lecture Notes In Computer Science, Vol. 1786, pp. 117- 131.

Smith, K.A., Gupta, J.N.D., 2002. Neural networks in business: techniques and applications, IRM Press, London.

Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., Sommerfield, D., 2001. “Visualizing data mining models”, in: Fayyad, U. Et al. (Eds.), Information visualization in data mining and knowledge discovery, Morgan Kaufman.