ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LXI 324 Number 7, 2013
http://dx.doi.org/10.11118/actaun201361072893

KNOWLEDGE DISCOVERY ON CONSUMERS' BEHAVIOUR

Pavel Turčínek, Arnošt Motyčka

Received: April 10, 2013

Abstract

TURČÍNEK PAVEL, MOTYČKA ARNOŠT: *Knowledge discovery on consumers' behaviour.* Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 2013, LXI, No. 7, pp. 2893–2901

This paper summarizes results of the research project "Application of modern methods to data processing in the field of marketing research" which was solved at the Department of Informatics, Faculty of Business and Economics of Mendel University in Brno. The most of these results were presented at international conferences.
It describes the use of knowledge discovery techniques on data from marketing research of consumers' behaviour. The paper deals with issues of classification, cluster analysis, correlation and association rules.
For classification there were used various algorithms: multi-layer perceptron neural network, self-organizing (Kohonen's) maps, bayesian networks and generation of a decision tree. Beside Kohonen's maps, which were tested in MATLAB software, all classification methods were tested in Weka software.
In order to find clusters of the methods K-means, Expectation-Maximization, DBSCAN Weka was also used as software for clustering.
Correlation analysis was done based on statistical approach. Generation of association rules was achieved by use of Apriori and the FP-growth algorithm in Weka.
The paper describes above mentioned methods and shows achieved results of exploring data from marketing research on consumers' behaviour.
This article discusses the suitability of these methods usage on such data sets. It also suggests further research possibilities of knowledge discovery on consumers' behaviour.

knowledge discovery, classification, cluster analysis, correlation, association rules, consumer behaviour, marketing research

The issue of consumer behaviour is explored in the field of customer relationship management. Customer Relationship Management (CRM) is seen as a holistic framework for interaction of organizations with their customers (Dařena, 2008). CRM uses marketing research as a strong tool to explore consumer behaviour. Marketing research is a process of collecting and using information for marketing decision making (Boone, Kurt, 2013) and plays an essential role in customer relationship management. Tools to facilitate individual steps of marketing research, particularly collection of data and their analysis can be more effective through increased use of databases and data mining techniques (Bradly, 2007; Kříž, Dostál, 2010). As a part of a Marketing Information System (Dařena, 2007) such tools provide decision makers with a continuous flow of information relevant to their area of responsibility (Boone, Kurt, 2013).

Every organization should focus on optimizing the workflows while ensuring compliance with regulation and dynamically responding to the market situation and customer requirements (Rábová, 2012). Use of Business Intelligence (Kříž, Klčová, Sodomka, 2011) and other modern approaches (Knížek *et al.*, 2011) are absolutely inherent in managerial decision-making in these days.

The aim of this paper is to summarize results of the research project "Application of modern methods to data processing in the field of marketing research" which was solved at the Department of

Informatics, Faculty of Business and Economics of Mendel University in Brno. Most of the results were presented at international conferences.

METHODS AND RESOURCES

The whole article focuses on exploring data from marketing research on customer behaviour in the Czech food market. The data file, on which the knowledge discovery is performed, has been acquired in the survey of the Department of Marketing and Trade of Faculty of Business and Economics, Mendel University in Brno. The questionnaire contained thirty items related to the research questions (low price, product composition, etc.), which respondents rated on scale from 1 to 10, where the value of 10 determined that this criterion had the highest importance to the interviewee. Then the other eight questions characterized the respondent (age, sex, educational level, etc.) (Turčínková, Kalábová, 2011).

Classification

The first attempt to discover some knowledge in this piece of data was the use of artificial intelligence methods (Miehie, Spiegelhalter, Taylor, 1994; Kohonen, Schroeder, Huang, 2001; Balogh, Klimeš, 2010; Weinlichová, Fejfar, 2010).

Multi-Layer Perceptron neural network (MLP)

Multi-Layer Perceptron neural network (MLP) is an acyclic forward network. Neurons can be divided into disjunctive layers so that the output of each neuron of one layer is connected to the inputs of each following neuron layer. There are no links between non-neighboring layers of neurons or between neurons in the same layer. Each neuron has as many inputs as there are neurons in the lower layer. The input layer serves only to distribute input values to the first hidden layer. A network with one hidden layer and one output layer is known as a two-layer network, a network with two hidden layers of a three-layer, etc. (Miehie, Spiegelhalter, 1994; Sarle, 1994). As a learning algorithm for the MLP neural network is the most widely used the back-propagation algorithm.

Back propagation algorithm is an iterative method where the network gets from an initial non-learned state to the full learned one (Miehie, Spiegelhalter, 1994; Sarle, 1994; Škorpil, Šťastný, 2008).

Self-organizing (Kohonen's) maps

Kohonen network is one of the networks that do not need to be trained by teacher. The basic idea of its function is cluster analysis that means the network ability to find certain characteristics and dependencies in the presented training data with the absence of any outside information (Kohonen, Schroeder, Huang, 2001; Kohonen *et al*., 1996; Fejfar, Šťastný, 2011).

In the Kohonen network is besides the input layer the output layer only. Number of inputs that come into the neuron is equal to the number of inputs to the Kohonen network. The weights of these inputs are used to encode the patterns which represent presented patterns. These neurons do not have their own transfer function. The only operation performed by the neuron is the calculation of distances (deviations) from the model presented pattern encoded in the weights of the neuron according to the relation (Kohonen, Schroeder, Huang, 2001; Kohonen *et al*., 1996):

$$d = \sum_{i=0}^{N-1} [x_i(t) - w_i(t)]^2 , \quad (1)$$

where $x_i(t)$ are the individual elements of the input pattern and $w_i(t)$ corresponding neuron weights, which are encoded patterns. Closer information can be found in (Kohonen, Schroeder, Huang, 2001; Kohonen *et al*., 1996).

Bayesian Networks

Bayesian classifiers are statistical classifiers. They can predict class membership probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes theorem described in (Han, Kamber, 2006).

Learning Bayesian networks from data has been known for a long time. This is a form of unsupervised learning, in the sense that the learner does not distinguish the class variable from the attribute variables in the data. The objective is to induce a network (or a set of networks) that "describes best" the probability of distribution over the training data. This optimization process is implemented in practice by using heuristic search techniques to find the best candidate over the space of possible networks. The search process relies on a scoring function that assesses the merits of each candidate network (Friedman, Geiger, Goldszmidt, 1997).

Decision tree

Decision trees are a way to represent rules underlying data with hierarchical, sequential structures that recursively partition the data. A decision tree can be used for data exploration in one or more of the following ways:

- Description: To reduce a volume of data by transforming it into a more compact form which preserves the essential characteristics and provides an accurate summary.
- Classification: Discovering whether the data contains well-separated classes of objects, so that the classes can be interpreted meaningfully in the context of a substantive theory.
- Generalization: Uncovering a mapping from independent to dependent variables that is useful for predicting the value of the dependent variable in the future (Murthy, 1998).

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision

tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The basic algorithm is described in (Han, Kamber, 2006).

Cluster analysis

Cluster analysis is a multidimensional statistical method that is used to classify objects. The basic problem of cluster analysis is to classify objects into groups (clusters) so that the two objects in the same cluster are more similar than two objects of different clusters (Řezanková, Húsek, Snášel, 2007).

The first problem is to determine the similarity of two objects. To be measured in similarity, each object must be characterized by their properties (Řezanková, Húsek, Snášel, 2007). Properties of objects can be divided into several categories. Han and Kamber (2006) reported these types of variables:

- Interval-Scaled variables
- Binary variables
- Categorical variables
- Ordinal variables
- Ratio-Scaled variables.

How to determine the similarity, respectively differences of individual criteria is described in detail (Han, Kamber, 2006; Grabmeier, Rudolph, 2002) and others.

You can find very many algorithms for creating clusters. It is however difficult to divide them clearly into different categories as some algorithm of them may exceed its category. Han and Kamber (2006) offer the following breakdown:

- Partitioning methods
- Hierarchical methods
- Density-based
- Grid-based
- Model-based
- Methods for clustering high-dimensional data.

Individual examples of algorithms are presented and described in (Řezanková, Húsek, Snášel, 2007; Han, Kamber, 2006; Grabmeier, Rudolph, 2002; Kaufman, Rousseeuw, 2005; Romesburg, 2004; Everitt, Landau, Leese, 2001; Ester, Kriegel, Sander, Xu, 1996).

Correlation analysis

In the most general sense, the word "correlation" refers to the degree level of association of two variables. Two variables are correlated (associated) if certain values of one variable tend to occur together with certain values of the other quantity.

The rate of association of two random variables can range from a lack of correlation to the absolute correlation. For a quantitative expression tightness of the relationship between two correlated variables exist a number of factors, which vary according to the types of variables for which they are used. The correlation between two continuous random variables X and Y is the most important and most frequently used when measured rate intensity of the relationship **Pearson correlation coefficient "*r*"**.

$$r = \frac{\sum\left[(x_1 - \bar{x}) \times (y_1 - \bar{y})\right]}{\sqrt{\sum (x_1 - \bar{x})^2} \times \sum (y_1 - \bar{y})^2}. \quad (2)$$

If we want to know exactly whether the correlation relationship really exists in the population, it is necessary to sample correlation coefficient **"r"**, as each selection parameter tested (Draper, Smith, 1998; Meloun, Militký, 2006).

Association rules

If we think of the universe as a set of items at the store, then each item has a Boolean variable representing the presence of that item. Each basket can then be represented by Boolean vector of values assigned to these variables. The Boolean vector can be analysed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in form of association rules. Example of buying computers and antivirus together is shown below:

Computer=>antivirus [support=2%, confidence=60%].

As you can see, there are two metrics how to measure rule interestingness. A **support** of 2% for Association Rule means that 2% of all transaction under analysis shows that computer and antivirus are purchased together. A **confidence** of 60% means that 60% of customers who purchased a computer also bought the antivirus software (Han, Kamber, 2006).

Apriori

The essential idea is to iteratively generate the set of candidate patterns of length ($k + 1$) from the set of frequent patterns of length k (for $k \geq 1$), and checks their corresponding occurrence frequencies (Han, Pei, Yin, Mao, 2004). More about Apriori algorithm can be found in (Han, Kamber, 2006; Han, Pei, Yin, Mao, 2004; Zaki, 2000).

Frequent pattern Mining

Frequent pattern mining plays an essential role in mining association rules. Most of the studies adopt an *Apriori*-like approach.

An interesting method is **frequent-pattern growth** (**FP-growth**), which adopts a *divide-and-conquer* strategy (Han, Kamber, 2006).

First, the compact data structure, called frequent pattern tree (FP-tree), is constructed, which is an extended prefix-tree structure storing crucial, quantitative information about frequent patterns. Only frequent length-1 items will have nodes in the tree.

Second, an FP-tree-based pattern fragment growth mining method, is developed, which starts from a frequent length-1 pattern, examines only its conditional pattern base, constructs its (conditional)

FP-tree, and performs mining recursively with such a tree. The pattern growth is achieved via concatenation of the suffix pattern with the new ones generated from a conditional FP-tree.

Third, the search technique employed in mining is a partitioning-based, divide-and-conquer method rather than Apriori-like bottom-up generation of frequent item sets combinations. This dramatically reduces the size of conditional pattern base generated at the subsequent level of search as well as the size of its corresponding conditional FP-tree (Han, Pei, Yin, Mao, 2004).

Results

In our research we mostly use the software Weka. Weka (Waikato Environment for Knowledge Analysis) is a machine learning tool in Java written, developed at the University of Waikato, New Zealand. WEKA is freely available software under the GNU General Public License.

Weka is a set of machine learning algorithms designed for data mining tasks. Algorithms can be applied directly to a data file, or you can call them via our own code written in Java. Weka contains tools for preprocessing, classification, correlation, clustering, association rules and visualization. It is suitable also for developing new machine learning schemes (Weka, 2012).

Classification

Our first goal was to test whether it is possible to classify the characteristics of respondents based on their answers to questions about their consumer behavior in the food market. In Šťastný, Turčínek, Motyčka (2011) is fully described the use of MLP and Kohonen's maps. In this paper we show just these results.

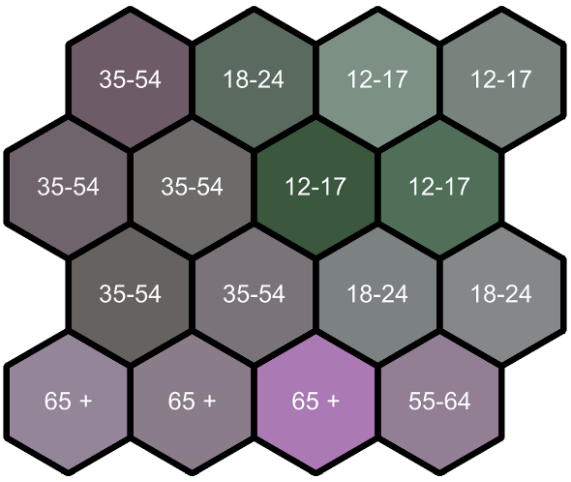
Tab. I shows the result of classification with use of the MLP algorithm. In this case we tried to classify the respondents into right age group according to their answering questions about their consume behaviour.

When we calculate the success rate of the correct classification, we find that it is only 44.5%.

The second approach to data mentioned above was the use of self-organizing (Kohonen's) maps. For the calculation there was used MATLAB with the use of Neural Network toolbox (Kohonen, 1996; Fejfar, Šťastný, 2011). Here we tried to train the network so that coalesced according to the characteristics of individual respondents. Results can be seen in Fig. 1.

Coloring principle of individual components lies in the fact that the richer the color is, the more similarity between the elements contained in this section it represents.

As it is clear from Figure 1 application of Kohonen's maps we have already managed to find a satisfactory way to group individual respondents' answers to the corresponding age categories logically grouped.



1: *Result by age (Šťastný, Turčínek, Motyčka, 2011)*

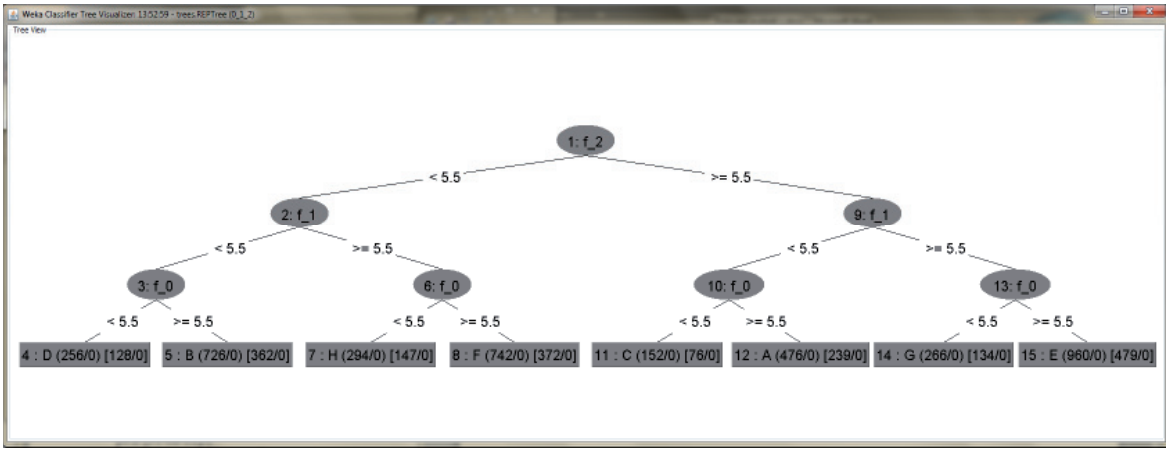
The second attempt, described in Turčínek, Šťastný, Motyčka (2012a), was to classify these items into groups which were stated by workers of Department of Marketing and Trade. These eight groups were based on three factors. We tried to find a way to classify these items into the same group based on different factors. For this purpose we used three various algorithms: MLP, Bayesian Networks, REPTree (one of Weka (2012) algorithms for a decision tree).

At first we classified based on the same factors and all algorithms classified without any mistake. The Fig. 2 shows classification by the REPTree algorithm.

In creating models for classification based on a combination of other factors we focused on the percentage of correctly classified instances and the time needed to build the model. Overview of the

I: *Inclusion in the classification of age (Šťastný, Turčínek, Motyčka, 2011)*

		classification					
		18–24	**35–54**	**25–34**	**12–17**	**65+**	**55–64**
reality	18–24	428	110	97	38	18	26
	35–54	116	334	88	6	36	28
	25–34	114	132	39	15	12	5
	12–17	52	13	8	30	2	4
	65 +	30	49	14	3	90	17
	55–64	29	55	18	2	22	10



2: *Decision tree generated based on the original factors by Weka (Turčínek, Šťastný, Motyčka, 2012a)*

II: *Summary results achieved by individual methods (Turčínek, Šťastný, Motyčka, 2012a)*

classification method	average classification success [%]	average time to build the model[s]	number of cases where the method was the best	the best case
BayesNET	34.2344	0.0212	109	57.8033
Multilayer Perceptron	33.8044	10.2395	40	56.9517
REPTree	33.0077	0.0357	5	57.3166

results can be found in Tab. II. This table is based on 154 combinations of input factors. It does not comprise in the original triplet.

The Tab. II shows that the success in classifying into classes based on the three input factors (second column) was very low. All created models showed that the original classification is highly dependent on three factors mentioned above. More detail can be found in Turčínek, Šťastný, Motyčka (2012a).

Cluster analysis

In finding clusters we have used several algorithms: K-means algorithm, Expectation-Maximization, DBSCAN and the algorithm for hierarchical clustering. Because of absence of clusters number knowledge in the dataset, at first we focused on the methods which do not require this information. As input criteria there were selected thirty items related to the issue of consumer behavior.

The first tested method was DBSCAN, however this did not give us reasonable results. As the second division option hierarchical clustering was chosen. Even using this method, we have not come to the desired distribution.

Another method which has been tested was Expectation-Maximization. When gaining the outputs of this method there were gradually

III: *Results of EM method (Turčínek, Šťastný, Motyčka, 2012b)*

No.	minStdDev	number of iterations	Number of clusters (the frequency of individual clusters)
1	0.000001	100	**10** (166, 251, 234, 267, 176, 195, 138, 274, 156, 163)
2	0.000001	200	**10** (166, 251, 234, 267, 176, 195, 138, 274, 156, 163)
3	0.00001	100	**10** (166, 251, 234, 267, 176, 195, 138, 274, 156, 163)
4	0.0001	100	**10** (166, 251, 234, 267, 176, 195, 138, 274, 156, 163)
5	0.0001	500	**10** (166, 251, 234, 267, 176, 195, 138, 274, 156, 163)
6	0.001	100	**10** (166, 251, 234, 267, 176, 195, 138, 274, 156, 163)
7	0.005	1000	**10** (166, 251, 234, 267, 176, 195, 138, 274, 156, 163)
8	0.01	100	**7** (472, 342, 237, 189, 323, 282, 175)
9	0.1	100	**7** (472, 342, 237, 189, 323, 282, 175)
10	0.5	100	**8** (308, 305, 151, 288, 290, 256, 249, 173)
11	1	100	**10** (178, 244, 193, 356, 99, 165, 132, 293, 206, 154)
12	1	1000	**10** (176, 244, 191, 358, 101, 164, 133, 292, 206, 155)
13	10	100	**11** (192, 211, 72, 234, 108, 159, 162, 228, 137, 162, 355)
14	100	100	**11** (192, 211, 72, 234, 108, 159, 162, 228, 137, 162, 355)

adjusted values of the minimum standard deviation (*minStdDev*) and the maximum number of iterations. Tab. III demonstrates the results.

As the Tab. III shows, the number of iterations does not influence the result too much. Besides one case where the number of elements in different clusters differed by a maximum of two, there was of no effect the number of iterations.

From the beginning of the experiments, it seemed that even change of the minimum standard deviation does not change the number and composition of clusters. The change occurred between the values 0.005, 0.01 where the number of clusters decreased from ten to seven. Another increase in this parameter brought then an increase of the number of clusters again.

This method of identifying clusters has provided results which, at first glance appear to be applicable. The output is a few clusters with an acceptable number of objects.

As the last method the *k*-means algorithm was used. This procedure requires the knowledge of the clusters number. We have achieved clusters of comparable size with this method. However, it is difficult to determine the number closest to reality.

More details about clustering of these data items can be found in Turčínek, Šťastný, Motyčka (2012b).

Correlation analysis

Our first step was to find any correlation between the individual answers. We compared all thirty answers one to each other and looked for cases where Pearson correlation coefficient *r* would be higher than 0.4 and significant according the earlier mentioned test. Tab. IV shows the numbers of pairs with certain *r* coefficient.

IV: *Numbers of pairs according Pearson correlation coefficient*

Pearson correlation coefficient	Numbers of pairs
0.40–0.49	9
0.50–0.59	3
0.60–0.69	1
0.70–0.79	0
0.80–0.89	0
0.90–1.00	1

The highest correlations (***r* = 0.9162**.) was between questions

- *I prefer Czech food to foreign food.*
- *If there is a choice, I prefer local food (food typical of the region I live in).*

These results do not bring anything new but prove that the questionnaire was filled in a meaningful way. More details in Turčínek, Šťastný, Motyčka (2012c).

Association rules

At first we used the Apriori algorithm and looked for rules with confidence 75% and higher and the minimum support was set to 10%. Each question may be in one of five states. Dependencies between states are not taken into account.

This setting gave us 159 association rules. The highest confidence (95%) with support of 10% was:

I prefer Czech food to foreign food = 3 + *Czech origin* = = 5 => *If there is a choice, I prefer local food (food typical of the region I live in)* = 3.

This rule does not show anything unexpected but it proves that this algorithm works in a good way. In most of the rules (128) there are included two statements:

- *I prefer Czech food to foreign food*
- *If there is a choice, I prefer local food (food typical of the region I live in*).

These statements were always on different sides of those rules. There were other two pairs which were more frequently together, exactly eight times:

- *The product is a special offer* and *Low price*
- *I always prefer to shop where I can buy fresh food* and *Expiration date.*

All of these rules make sense and point out interesting information. Not very predictable connection is between statements *Czech origin* and *Expiration date*, which appears in five rules. In one of them there are just these two statements:

Czech origin = 5 => Expiration date = 5
[support=20%, confidence=77%].

This means that if customer highly cares about Czech origin, the expiration date is also very important to him/her.

There are more different rules however it is up to the marketing specialist whether they bring something useful.

We used the FP-growth algorithm as the second one for finding association rules. The use of this algorithm did not bring as interesting results as the first approach. It was also caused because of necessary data modification for this algorithm. All results and details can be found in Turčínek, Šťastný, Motyčka (2012c).

DISCUSSION

The whole research was based on the same data source (Turčínková, Kalábová, 2011). This brings us a view of exploring this data from different angles.

The use of classification showed that these pieces of data are hard to classify. Only using Kohonen's maps seemed to be suitable. The other methods were not able to classify into chosen groups, however it does not mean that they would not work for different groups or different settings.

In this research cluster analysis of the above mentioned data was performed by the following

methods: DBSCAN, HierarchicalClusterer, Expectation-Maximization, K-means. The analysis shows that for a given data set there are the only suitable methods EM and K-means, which created useable (reasonable) clusters out of input data. These methods can be used with advantage in the preparatory phase of the subsequent application of the methods for dealing with data classification according to the set parameters.

The found correlations among individual characteristics proved that questionnaires were filled meaningfully because information we got makes sense. Unfortunately we did not find any unexpected relation among characteristics.

The use of the Apriori algorithm for finding association rules generated the same expected rules, however it brought also some new information such as connection between Czech origin and expiration date. All generated rules will be discussed with marketing specialists.

The use of the FP-growth algorithm did not bring good results. It was caused by transformation of the data. The transformation made the data too unspecific so it was hard to set the settings to find interesting information.

The research will continue with a different set of data. The future work will be focused on new methods for regression that represent mainly evolutionary algorithms. Most of the methods found in statistical analysis are not eligible since most of the real-world problems have nonlinear character (Štencl, Popelka, Šťastný, 2011). Linear regression might be used only on short intervals of the measured data or under restricted conditions. These tasks are therefore suitable for their solving using genetic algorithms and other evolutionary methods (Popelka, Šťastný, 2007).

It's possible to solve the regression problem by means of a grammatical evolution method. Grammatical evolution is a method based on a genetic algorithm extended with compiler of context-free rewriting grammar and other supplementary algorithms (Popelka, Šťastný, 2007). The output of a grammatical evolution method is a solution of the given problem in a symbolic form. This is very useful especially for non-linear regression problems. The output is then a standard math function not bound to the optimization software and it is possible to use it in other applications (Popelka, Šťastný, 2011).

SUMMARY

The whole paper is a summarization of results of the research project "Application of modern methods to data processing in the field of marketing research" which was solved at the Department of Informatics, Faculty of Business and Economics of Mendel University in Brno. The vast majority of these results were presented at international conferences.

The article describes the use of knowledge discovery techniques on data from marketing research of consumers' behaviour. Paper focused on issues of classification, cluster analysis, correlation and association rules.

For classification there were used several algorithms such as multi-layer perceptron neural network, self-organizing (Kohonen's) maps, bayesian networks and generation of a decision tree. Beside Kohonen's maps, which were tested in MATLAB software, all other classification methods were tested in Weka software.

For cluster analysis were also used software Weka. There were used of methods K-means, Expectation-Maximization, DBSCAN and others.

Correlation analysis was done based on statistical approach. Generation of association rules was achieved by use of Apriori and the FP-growth algorithm in Weka.

The article describes above mentioned methods and shows achieved results of exploring data from marketing research on consumers' behaviour.

This paper discusses the suitability of used methods usage on such data sets. It also suggests further research possibilities of knowledge discovery on consumers' behaviour.

Acknowledgement

This work has been supported by the research design of Mendel University in Brno MSM 6215648904/03.

REFERENCES

BALOGH, Z., KLIMEŠ, C., 2010: Modelling of education process in LMS using Petri nets structure. In: *Proceedings of the IADIS International Conference e-Learning 2010, Part of the IADIS Multi Conference on Computer Science and Information Systems 2010, MCCSIS 2010*. Freiburg; Germany: *IADIS*, 2: 289–291. ISBN 978-972893917-5

BOONE, L. E., KURT, D. L., 2013: *Contemporary Marketing*. Mason: Cengage Learning, 766 p. ISBN 978-1-111-57971-5.

BRADLEY, N. *Marketing Research: Tools and Techniques*. New York: Oxford University Press, 531 p. ISBN 0-19-928196-3.

DAŘENA, F., 2008: A research on CRM systems in the Czech Republic. *Acta Univ. Agric. et Silvic. Mendel. Brun.*, 56, 3: 29–34. ISSN 1211-8516.

DAŘENA, F., 2007: Global architecture of marketing information systems. *Agricultural Economics*, 52, 9: 432–440. ISSN 0139-570X.

DRAPER, N. R., SMITH, H., 1998: *Applied Regression Analysis*. 3rd ed. New York: Wiley, 706 p. ISBN 978-0-471-17082-2.

ESTER, M., KRIEGEL, H., SANDER, J., XU, X., 1996: A density-based algorithn for discovering clusters in large databases with noise. *Proceedings of the 2nd ACM SIGKDD International Conference on Knoelwdge Discovery and Data Mining*. Portland: AAAI Press, p. 226–231. on-line: http://dns2.icar.cnr.it/manco/Teaching/2005/datamining/articoli/KDD-96.final.frame.pdf. [cit. 2013-02-21].

EVERITT, B. S., LANDAU, S., LEESE, M., 2001: *Cluster Analysis*. 4th ed., London: Arnold, 237 p., ISBN 978-0-340-76119-9.

FEJFAR, J., ŠŤASTNÝ, J., 2011: Time series clustering in large data sets. *Acta Univ. Agric. et Silvic. Mendel. Brun.*, 59, 2: p. 75–80. ISSN 1211-8516.

FRIEDMAN, N., GEIGER, D., GOLDSZMIDT, M., 1997: Bayesian Network Classifiers. *Machine Learning*, 29, 2: 131–163. ISSN 0885-6125.

GRABMEIER, J., RUDOLPH, A., 2002: Techniques of Cluster Algorithms in Data Mining. *Proceedings of Data Mining and Knowledge Discovery*, 6, 4: 303–360. ISSN 1573-756X.

HAN, J., KAMBER, M., 2006: *Data Mining: Concepts and Techniques*. 2nd ed. San Francisco: Morgen Kaufmann Publishers, 770 p. ISBN 978-1-55860-901-3.

HAN, J., PEI, J., YIN, Y., MAO, R., 2004: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8, 1: 53–87. ISSN 1384-5810.

KAUFMAN, L., ROUSSEEUW, P. J., 2005: *Finding Groups in Data: An Introduction to Cluster Analysis*. 2nd ed., New Jersey: John Wiley & Sons, 342 p. ISBN 0-471-73578-7.

KNÍŽEK, J., ŠINDELÁŘ, J., VOJTĚŠEK, B., BOUCHAL, P., NENUTIL, R., BERÁNEK, L., DĚDÍK, O., 2011: Using markers to aid decision making in diagnostics. *International Journal of Tomography and Statistics*, 16, W11: 41–55. ISSN 0972-9976.

KOHONEN, T. et al., 2012: *SOM PAK: The Self-Organizing Map Program Package*. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, 1996. on-line: http://www.cis.hut.fi/research/papers/som_tr96.ps. [cit. 2012-10-11].

KOHONEN, T. SCHROEDER, M. R., HUANG, T. S., 2001: *Self-Organizing Maps*. New York: Springer-Verlag, Inc., 511 p. ISBN 3540679219.

KŘÍŽ, J., DOSTÁL, P., 2010: Database System and Soft Computing. *Systémová integrace*, 17, 4: 17–26. ISSN 1210-9479.

KŘÍŽ, J., KLČOVÁ, H., SODOMKA, P., 2011: The Use of Business Intelligence Tools for Prediction and Decisionmaking Processes in the Academic Environment: A Case Study. *Innoviation and Knowledge Management – A Global Competitive Advantage, Proceedings of The 16th International Business Information Management Association Conference. IBIMA*. Kuala Lumpur, Malaysia: International Business Information Management Association (IBIMA), p. 1259–1265. ISBN 978-0-9821489-5-2.

MELOUN, M., MILITKY, J., 2006: *Kompendium statistického zpracování dat: metody a řešené úlohy*. 2nd ed. Praha: Academia, 982 p. ISBN 80-200-1396-2.

MIEHIE, D., SPIGELHALTER, D. J., TAYLOR, C. C., 1994: *Machine Learning, Neural and Statistical Classification*. New York: Horwood Limited, Ellis, 289 p. ISBN 013106360X

MURTHY, S. K., 1998: Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2, 4: p. 345–389. ISSN 1384-5810.

POPELKA, O., ŠŤASTNÝ, J., 2011: Automatic Generation of Programs. In: Dr. MATTHIAS SCHMIDT (Ed.), *Advances in Computer Science and Engineering*, ISBN 978-953-307-173-2, InTech, DOI: 10.5772/16012. Available from: http://www.intechopen.com/books/advances-in-computer-science-and-engineering/automatic-generation-of-programs.

POPELKA, O., ŠŤASTNÝ, J., 2007: Generation of mathematic models for enviromental data analysis. *Management si Inginerie Economica*, 23, 6: 61–66. ISSN 1583-624X.

RÁBOVÁ, I., 2012: Using UML and Petri nets for visualization of business document flow. *Acta Univ. Agric. et Silvic. Mendel. Brun.*, 60, 2: 299–306. ISSN 1211-8516.

ROMESBURG, H. C., 2004: *Cluster Analysis For Researchers*. North Carolina: Lulu Press, 334 p., ISBN 1-4116-0617-5.

ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V., 2007: *Shluková analýza dat*. Praha: Professional Publishing, 196 p. ISBN 978-80-86946-26-9.

SARLE, W. S., 1994: Neural Networks and Statistical Models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference, Cary*, NC: SAS Institute, p. 1538–1550. on-line: ftp://ftp.sas.com/pub/neural/neural1.ps. [cit. 2013-02-11].

ŠKORPIL, V., ŠŤASTNÝ, J., 2008: Comparison of Learning Algorithms. *24th Biennial Symposium on Communications*. Canada: Kingston, p. 231–234. ISBN 978-1-4244-1945-6.

ŠTENCL, M., POPELKA, O., ŠŤASTNÝ, J., 2011: Comparison of time series forecasting with artificial neural network and statistical approach. *Acta Univ. Agric. et Silvic. Mendel. Brun.*, 59, 2: 347–352. ISSN 1211-8516.

ŠŤASTNÝ, J., TURČÍNEK, P., MOTYČKA, A., 2011: Using Neural Networks for Marketing Research Data Classification. *International WSEAS Conference on Mathematical Methods and Techniques in Engeneering & Environmental Science*. Catania, Italy: WSEAS Press, 2p. 252–256. ISBN 978-1-61804-046-6.

TURČÍNEK, P., ŠŤASTNÝ, J., MOTYČKA, A., 2012a: Usage of Data Mining Techniques on Marketing

Research Data. *Proceedings of the 11th WSEAS International Conference on Applied Computer and Computational Science (ACACOS '12)*. Rovaniemi, Finland: WSEAS Press, p. 159–164. ISBN 978-1-61804-084-8.

TURČÍNEK, P., ŠŤASTNÝ, J., MOTYČKA, A., 2012b: Usage of cluster analysis in consumer behavior research. *Proceedings of the 12th WSEAS International Conference on APPLIED INFORMATICS AND COMMUNICATIONS (AIC '12)*. Istanbul, Turkey: WSEAS Press, p. 172–177. ISBN 978-1-61804-113-5.

TURČÍNEK, P., ŠŤASTNÝ, J., MOTYČKA, A., 2012c: Finding Dependencies in Food Market Consumer Behavior. *Proceedings of the European Conference of COMPUTER SCIENCE (ECCS '12).* Paris, France: WSEAS Press, p. 67–72. ISBN 978-1-61804-140-1.

TURČÍNKOVÁ, J., KALÁBOVÁ, J., 2011: Preferences of Moravian consumers when buying food. *Acta Univ. Agric. et Silvic. Mendel. Brun.,* 59, 2: 371–376. ISSN 1211-8516.

WEINLICHOVÁ, J., FEJFAR, J., Usage of self-organizing neural networks in evaluation of consumer behaviour. *Acta Univ. Agric. et Silvic. Mendel. Brun.,* 58, 6: p. 625–632. ISSN 1211-8516.

WEKA, 2012: *Weka 3 – Data Mining with Open Source Machine Learning Software in Java*. [on-line]. HTML Document. 2011. [cit. 2012-10-11]. http://www.cs.waikato.ac.nz/ml/weka/.

ZAKI, M. J.,2000: Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12, 3: 372–390. ISSN 1041-4347.

Address

Ing. Pavel Turčínek, doc. Ing. Arnošt Motyčka, CSc., Department of Informatics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic, e-mail: pavel.turcinek@mendelu.cz, mot@mendelu.cz