ANALYZING CONSUMER BEHAVIOR USING NEURAL NETWORKS AND GRAMMATICAL EVOLUTION

Aleš Ďurčanský[1], Kamil Staněk[2], Michal Ježek[2], Jiří Šťastný[1,2]

[1]Mendel University in Brno, Zemedelska 1665/1, 613 00 Brno, Czech Republic
[2]Brno University of Technology, Technicka 2896/2, 619 69 Brno, Czech Republic

ABSTRACT

In this contribution, we will present an approach to the automatic classification of customers based on their behaviour in the food market. The analysis is based on the data from a survey on meat product consumption in the Czech Republic. Classifiers were created to categorize customers into classes according to their habits of purchasing meat products, dividing the customers with respect to such characteristics as age or education. To accomplish this task some selected types of artificial neural networks (Multi-Layer Perceptron Neural Network, Kohonen Neural Network) were trained and also an approach based on grammatical evolution was used. These classifiers were compared with regard to their abilities to perform the given task. Also, the survey data pre-processing is described.

Keywords: Consumer's Behaviour, survey analysis, classification, grammatical evolution, neural networks

1 INTRODUCTION

All organizations should pay attention to optimizing their workflows, also follow regulations, and dynamically respond to the situation on the market and customer needs (Rábová, 2012). Customer Relationship Management (CRM) can be viewed as a holistic framework that allows interaction between organizations and their customers (Dařena, 2008). To explore consumer behaviour, CRM uses marketing research. Marketing research is a process of collecting and using information for marketing decision-making (Boone and Kurt, 2013).

Birčiaková *et al.* (2014) describe consumer society by many distinctive features: increasing consumer activities, the new phenomenon of recreational shopping, strongly location-based consumption, strengthening customers' role on the market, developing IT and its impact on consumer behaviour in the form of a broader and more varied selection and availability of products and services, and easier access to the information from both the supply and demand side. There are a lot of factors that influence customer behaviour (Hajko *et al.*, 2014). Factors which influence the behaviour of consumers are very important for businesses because they

https://doi.org/10.11118/978-80-7701-047-4-0051

can focus clearly on their business policy based on these factors, which should lead to better business results (Novotný and Duspiva, 2014). The knowledge of fundamental relations gives the possibility to realize predictions and helps with the decide-making process about reasonable actions in order to achieve the desired objectives (Bína and Jiroušek, 2015).

Data for research on consumer behaviour can be obtained from multiple sources. In the secondary research being typically used in national and international sources, such as the Czech Statistical Office or Eurostat, data is provided in electronic form, easily accessible via the Web. In the primary research, the most often used data is from surveys in which consumers respond to specific questions.

Tools that enable individual steps of marketing research, particularly the collection of data and its analysis can be more effective through the increased use of databases and data mining techniques (Bradly, 2007; Kříž and Dostál, 2010; Munk *et al.*, 2013, Beránek and Knížek, 2012). As a part of a Marketing Information System, these tools provide persons with decision responsibility with an instant flow of information relevant to their area. (Boone and Kurt, 2013).

The aim of this contribution is the automatic classification of customers based on their behaviour in a food market. The analysis is based on the data from a survey on meat product consumption in the Czech Republic. This article is related to the research from Lýsek and Šťastný (2013) where grammatical evolution was tested for survey data classification and compared to neural networks from Šťastný *et al.* (2011).; other applications of these methods include customer segmentation tasks (e.g., Mitchell, 1999 for GA; Kohonen, 1982 for SOM; Skorpil and Stastny, 2006 for MLP comparisons)

This study addresses two key research questions (RQ):

- RQ1: Is it possible to classify consumer types using passively collected behavior data instead of direct survey responses?
- RQ2: Which classification method (MLP, Kohonen network, grammatical evolution) provides the best accuracy in this domain?

It is very costly and time-consuming to ensure a sufficiently high-quality prediction of customer behaviour. We must also take into account data protection and the willingness of customers to disclose their data. It is better to avoid these problems and find minimized easy-to-detect criteria for customer behaviour analysis. This paper shows multiple techniques that solve the above-described problem. The use of neural networks and genetic algorithms is also shown as a possible solution to these analyses.

2 MATERIALS AND METHODS

For analysing customer behaviour, it is quite common to use data from marketing survey research (Turcinkova, Stavkova, 2012; Litavcova *et al.*, 2015). The data used in our research was gathered in a customer survey on meat product consumption, the same data items were also used in Turcinkova *et al.* (2014). Customers answered a set of questions based on their opinions and preferences. There were also a few questions the purpose of which was to categorize each participant by such parameters as age, gender, education, type of household, size of his/her hometown or employment status, and a few others. The survey contained 1027 responses.

The survey responses were stored as 0 or 1 values for yes/no questions. The ranking questions stored the response as numbers. If the answer to a survey question is enumerative, the answer was stored as an index of the corresponding answer. We selected four parameters that the survey participants indicated (age, gender, education, and employment) as target classifications to test the proposed algorithms.

2.1 Selection of classification indices

Our goal was to create a classifier which would be able to recognize customers by the least number of parameters. The ideal state is that the customer is classified without even knowing about it – we do not want the customer to fill out any survey forms. The survey data items were therefore filtered out from the information that is undetectable without asking the customer directly.

Most customers use their credit cards or some kind of store membership/discount card.

We can track and connect the shopping sessions of these customers and gather information such as the frequency of shopping, the time when a customer visits the shop most, the amount of money spent, and the type of goods bought. This information was gathered from a survey but it can also be gathered from the company's accounts.

To satisfy our goal we selected the following customer parameters from this survey data:

- What days do you usually do your shopping for meat products? (8 options)
- What time do you usually go shopping for meat products on weekdays and what time on weekends? (10 options)
- How often do you usually buy meat products? (15 options)
- What type of payment do you usually use when shopping for meat products? (3 options)
- Frequency of purchasing selected meat products classes (68 options).

The resulting vector describing a customer had 104 indices. Table 1 shows classification criteria and their number of classes.

Tab. 1 Number of classes

Classification criteria	Number of classes
Age	6
Gender	2
Education	5
Employment status	8

2.2 Pre-processing of training data

To select a training set of data for classifiers, a 1NN search was executed. Only instances of survey responses with the same class label as their nearest neighbour were selected to enter the training set. The training set size was from 1/3 to 1/2 of the original dataset after this pre-processing step.

Response values for enumerative questions were converted into binary indicators by creating a new column for each possible response. That column contained the value 1 if the survey participant selected that option or the value 0 if he/she did not.

2.3 Used methods

Classifier configurations were selected based on prior experiments and related literature. No extensive hyperparameter tuning was conducted due to computational constraints. For each method, a simple train-test split (approximately 70/30) was used without stratification. This setup was intended to simulate realistic classification performance in a CRM context.

Since the Multi-Layer Perceptron (MLP) network is often used for traditional classification tasks (Skorpil and Stastny, 2006 and 2009), we will include a basic implementation of this network as one of proposed classifiers.

The Kohonen neural network (or Self-organizing map) by Kohonen (1982) will be used as the second type of the classifier. As for the given task we expect that it will perform adequately as the nature of the input data is a vector of indices and this network can naturally cluster such a kind of input (Konecny *et al.*, 2010).

Another type of the classifier will be created using genetic programming introduced by Koza (1992), namely the grammatical evolution approach will be used. The performance of these classifiers depends heavily on the given grammar which is used to build a short computer program. That program is used as the classifier in the final phase.

Grammatical evolution is an optimization process which can be used to develop short computer programs. It is an evolutionary algorithm. These algorithms are inspired by the process of the evolution in nature (Mitchell, 1999; Goldberg, 2002). These algorithms work in an iterative manner with a population of individuals where each individual represents a candidate solution to the problem (see Fig. 1).

Grammatical evolution uses context free grammar G = {N, T, P, S} to translate a sequence of numbers (chromosome) into a computer program (Ryan and O'Neill, 2003). The grammar consists of these components: N is a set of non-terminal symbols, T is a set of terminal symbols, P is a set of production rules and S is the starting symbol from N set (Hopcroft and Ullman, 1969). We used backward processing modification of chromosome translation in our grammatical evolution framework proposed by Popelka and Stastny (2011).

The grammar we used for this classification task is the same as in Lysek and Stastny (2013). The grammar can create classifier programs with multiple output values using the semicolon node, (Lysek and Stastny, 2019).

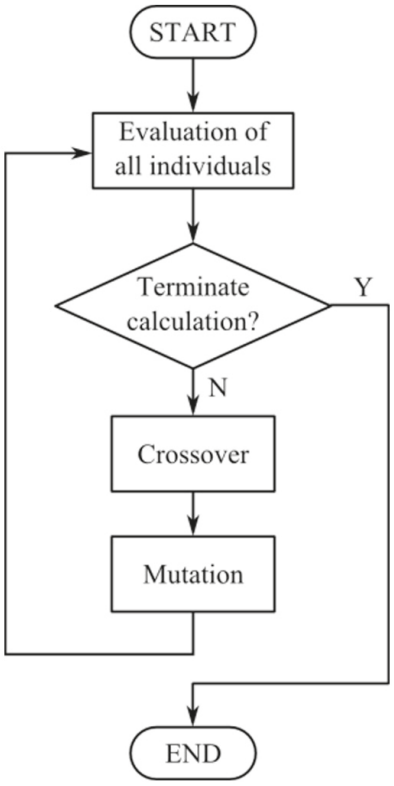

Fig. 1: Genetic algorithm visual representation

The purpose of the program created by grammatical evolution is not restricted. The process is guided by a fitness value which describes the performance of the evolved program for the given task. The fitness value of the member M is calculated according to the formula (1).

$$Fit(M) = W_1 SR + W_2 \frac{RC}{CC} \tag{1}$$

$$W_1 = \frac{CC}{CC+1} \tag{2}$$

$$W_2 = \frac{1}{CC+1} \tag{3}$$

$$SR = \frac{\sum_{i=1}^{CC} \frac{S_i}{MS_i}}{CC} \tag{4}$$

The given formula calculates the classification success rate in the first part. The SR value (4) is an average of success rates for each class i of CC classes. The second part encourages the classifier to be able to classify all classes as the variable selection of rewriting rules from the grammar allows an omission of some output values. Value RC is the number of classes that the classifier is capable of recognizing. These formulas were proposed for classification tasks in Lysek and Stastny (2019). Weights w1 and w2 distribute the importance of those two parts according to the number of classes.

3 RESULTS

We present the result and setup of selected artificial intelligence methods.

3.1 MLP network

A three layer (104 – 10 – N) MLP neural network was used. Back-propagation was used as the training algorithm for 100 iterations. The number of output neurons N was set according to the number of classes of the classification task. An output vector was M-dimensional for each task and each class was associated with one basis vector of M-dimensional space. Each basis vector was associated with one class. The classification result was a basis vector (it is an associated class label) closest to the neural network output for the given input. A Euclidean distance formula was used to calculate the distances of vectors. The result can be seen in the Table 2.

3.2 Kohonen network

The Kohonen network was trained on pre-processed training data. Two variants were tested. One with 64 output neurons and one with 25 output neurons. The training algorithm was executed for 200 iterations. The labelling of output neurons was done by measuring the most frequent class label for each neuron on training data. We used a square grid and corresponding

Tab. 2 Results of classification using MLP network with 25 output neurons

Results of classification	Training data performance [%]	Testing data performance [%]
Age	6.03	19.07
Gender	31.78	39.30
Education	54.83	41.08
Employment status	35.50	36.02

Tab. 3 Results of classification using Kohonen network with 25 output neurons

Results of classification	Training data performance [%]	Testing data performance [%]
Age	47.98	30.52
Gender	69.72	60.60
Education	59.35	42.05
Employment status	64.32	48.53

neighbourhood function for weight updates. A higher number of output neurons gives the network the ability to create a finer clustering and improves the classification quality. Tables 3 and 4 show the results.

3.3 Grammatical evolution

Result of grammatical evolution algorithm is shown in the Table 5.

The parameters of the grammatical evolution process are stated in the following list.

- Length of chromosome: 80.
- Number of iterations: 1000.
- Population size: 200.
- Mutation rate: 10%.
- Crossover rate: 90%.

Tab. 4 Results of classification using Kohonen network with 64 output neurons

Results of classification	Training data performance [%]	Testing data performance [%]
Age	58.62	31.76
Gender	72.73	62.46
Education	63.87	42.67
Employment status	70.15	52.17

Tab. 5 Results of classification using classifier evolved by grammatical evolution

Results of classification	Training data performance [%]	Testing data performance [%]
Age	37.93	17.83
Gender	56.02	54.30
Education	41.07	25.55
Employment status	50.60	33.62

4 DISCUSSION

The classification results are rather average. The reasons are many, the most significant reason would probably be that customers often act randomly and it is not really possible to gain better results from classification methods that work on the basis of finding patterns. Nevertheless, the presented Kohonen neural network is capable of quite a good classification performance on the given data.

The advantage of the Kohonen network is that it has the capability to create custom clusters by grouping similar input patterns. We can observe that with a rising number of output neurons the performance of classification grows as well. This fact indicates that the data contains more interesting information about the customer's patterns of behaviour but we do not have class labels for such a fine clustering.

The Grammatical evolution approach would perform better if we let the evolutionary process use longer chromosomes. The Classifier program using a chromosome of length 80 cannot be able to use all input terminals because the input vector has high dimensionality. On the other hand, a longer chromosome would increase the number of possible program forms and therefore the duration of the search for an optimal solution would increase significantly. We would prefer grammatical evolution for the classification of input vectors with significantly lower dimensionality.

Generally, the worst classification results are for the age classification and we think that age is not the main parameter that defines customer behaviour in the food market. In our opinion, the main parameter is the amount of money the customer is willing to spend and that is given by his/her education and employment status. Both of these characteristics showed a higher classification success rate. Also, the customer‘s gender can be determined with quite a high success rate.

5 CONCLUSION

A predictive customer behaviour analysis is a highly demanding activity not only financially but also in terms of time. Simplifying this analysis by using easily discoverable classifiers appears to be a possible solution.

By conducting research on classification methods using neural networks and genetic programming, their effectiveness in solving a consumer behaviour analysis is demonstrated. The research has shown that the methods were chosen correctly and thus the most appropriate classification method for this analysis can be derived. The use of the Kohonen network shows the most accurate results among the studied classification methods for broad-spectrum classifiers. The expectation of the best-performing solution when using the MLP network as a classification method was found to be insufficient in the setting used. However, even when using a Kohonen network, it is still very difficult to deliver an accurate analysis of customer behaviour approaching at least an 80 % success rate.

Although we performed the analysis correctly, we were limited by the performance of the specific classifiers used. Their independent potential does not allow for substantial progress in the analysis, but external influences, not mentioned in this paper, are the cause.

Further research on a better-performing solution should focus on refining the selection of specific classifiers. There also appears to be great potential in using a close correlation of at most two to three classifiers. The results of these analyses could be used, for example, to support managerial decision-making. From a practical perspective, the Kohonen network, due to its clustering capability and relatively robust performance, is a promising method for integration into CRM tools that aim to profile customers without direct engagement. This has implications for automated marketing strategies, loyalty systems, or dynamic pricing mechanisms.

REFERENCES

BERANEK, L., KNIZEK, J. 2012. The Usage of Contextual Discounting and Opposition in Determining the Trustfulness of Users in Online Auctions. *Journal of Theoretical and Applied Electronic Commerce Research*. 7(1), 34–50.

BINA, V., JIROUSEK, R. 2015. On computations with causal compositional models. *Kybernetika*. 51(3), 525–539.

BIRCIAKOVA, N., STAVKOVA, J., SOUCEK, M. 2014. How Marketing Instruments Affect Consumer Behavior in Times of Economic Turbulence. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. 62(6), 1257–1263.

BOONE, L. E., KURT, D. L. 2013. *Contemporary Marketing*. Mason: Cengage Learning.

BRADLEY, N. 2007. *Marketing Research: Tools and Techniques*. New York: Oxford University Press.

DARENA, F. 2008. A research on CRM systems in the Czech Republic. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 56(3), 29–34. https://doi.org/10.11118/actaun200856030029

GOLDBERG, D. E. 2002. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Boston: Kluwer Academic Publishers.

HAJKO, V., BIRCIAKOVA, N., STAVKOVA, J. 2014. *The trends in the income distributions in the EU-27 countries: Measuring the differences*. Paper presented at the 32nd International Conference Mathematical Methods in Economics.

HOPCROFT, J., ULLMAN, J. D. 1969. *Formal languages and their relation to automata*. Addison–Wesley.

KOHONEN, T. 1982. Self-Organized Formation of Topologically Correct Feature Maps.
Biological Cybernetics. 43, 59–69.

KONECNY, V., TRENZ, O., SVOBODOVA, E. 2010. Classification of companies with the assistance of self-learning neural networks. *Agricultural Economics*. 56(2), 51–58.

KOZA, J. R. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge MA: The MIT Press.

KRÍZ, J., DOSTAL, P. 2010. Database System and Soft Computing. *Systemova integrace*. 17(4), 17–26.

LITAVCOVA, E., BUCKI, R., STEFKO, R., SUCHANEK, P., JENCOVA, S. 2015. Consumer's Behaviour in East Slovakia after Euro Introduction during the Crisis. *Prague Economic Papers*. 24(3), 332–353. https://doi.org/10.18267/j.pep.522

LYSEK, J., STASTNY, J. 2013. Classification of Economic Data into Multiple Classes by Means of Evolutionary Methods. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 61(7), 2445–2449.

LYSEK, J., STASTNY, J. 2019. Grammatical evolution for classification into multiple classes. *Advances in Intelligent Systems and Computing*. 61(837), 192–207.

MITCHELL, M. 1999. *An Introduction to Genetic Algorithms*. Cambridge MA, MIT Press.

MUNK, M., DRLIK, M., KAPUSTA, J., MUNKOVA, D. 2013. Methodology design for data preparation in the process of discovering patterns of web users behaviour. *Applied Mathematics and Information Sciences*. 7(1), 27–36.

NOVOTNY, J., DUSPIVA, P. 2014. Factors Influencing Consumers' Buying Behavior and their Importance for Enterprises. *E+M Ekonomie a Management*. 17(1), 152–166. https://doi.org/10.15240/tul/001/2014-1-012.

POPELKA, O., STASTNY, J. 2011. Automatic generation of programs. In: *Advances in Computer Science and Engineering*. InTech Open, pp. 17–36. ISBN 978-953-307-173-2

RABOVA, I. 2012. Using UML and Petri nets for visualization of business document flow. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. 60(2), 299–306.

RYAN, C., O'NEILL, M. 2003. *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*. Kluwer.

SKORPIL, V., STASTNY, J. 2006. *Back-Propagation and K-Means Algorithms Comparison*. In: *2006 8th International Conference on Signal Processing*. Guilin, China. IEEE Press.

SKORPIL, V., STASTNY, J. 2009. *Comparison Methods for Object Recognition*. Paper presented at the 13th WSEAS International Conference on Systems. Rhodos, Greece.

STASTNY, J., TURCINEK, P., MOTYCKA, A. 2011. *Using Neural Networks for Marketing Research Data Classification*. Paper presented at the International WSEAS Conference on Mathematical Methods and Techniques in Engeneering & Environmental Science. Catania, Italy.

TURCINKOVÁ, J., STAVKOVA, J. 2012. Does the Attained Level of Education Affect the Income Situation of Households? *Procedia – Social and Behavioral Sciences. Special Issue 3rd International Conference on New Horizons in Education - INTE 2012*. 55, 1036–1042. https://doi.org/10.1016/j.sbspro.2012.09.595

TURCINKOVA, J., TURCINEK, P., MOTYCKA, A., STAVKOVA, J. 2014. *Exploring How Customers Shop for Meat Products*. Paper presented at the Recent Advances in Economics, Management and Marketing, Cambridge, MA, USA.

Acknowledgement

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, this is co-financed from Operational Programme Research, Development and Education.

Contact information

Aleš Ďurčanský: xdurcans@mendelu.cz,
Jiří Šťastný: stastny@fme.vutbr.cz