Classification of Clients on the basis of Modifying Case-based Reasoning Algorithms

Filip Mezera and Jiri Krupka

Institute of System Engineering and Informatics, Faculty of Economics and
Administration, University of Pardubice,
Studentska 84, 532 10 Pardubice, Czech Republic
mezera.filip@email.cz,jiri.krupka@upce.cz
http://www.upce.cz

Abstract. This article deals with client's classification possibilities for trade companies. One of the options for dealing with classification is to use Case-based Reasoning method. The proposed modified algorithm clears cases base, it eliminates invalid data and out-dated data from the time point of view. Hill-climbing algorithm is used for this. Classification results achieved by means of modified algorithm are compared with other classification methods such as Neuron Networks, Top Down Induction Decision Trees and Logistic Regression.

Keywords: Aging of cases, Base clearing, Case-based reasoning, Clients' classification, Customer relationship management

1 Introduction

Majority of private companies (hereinafter PCs) deal with a number of problems. One of such problems is classification of clients and definition of trade relation parameters (prices, margins, charges and similar). Correct definition of these parameters is essential for relations with customers, it motivates customers to remain in such business relations (better satisfied clients' needs keeping a certain level of profit). Clients make their decisions based on a number of factors [12]. These factors cannot be easily categorized. PCs usually use for this classification their own business data, other generally available data respectively, e.g. geographic data.

A classification model can be realized by means of a number of methods, from statistical methods (a cluster analysis) to computational intelligence methods (neural networks, fuzzy inference system etc.). One of possible approaches is also the use of case-based reasoning (CBR) [1], [3], [4], [6], [7], [14], [16], [19], [20].

The objective of this article is to design a classification model of modified CBR algorithm. In comparison with the classic CBR this model is extended by clearing of the case-basis from invalid cases and by time aspect of cases comparison (aging of cases in base). The designed algorithm is optimized with the objective of the highest exactness of classification by means of Hill-climbing method.

Achieved results of the classification model are compared with other classification methods by means of Top Down Induction Decision Trees (TDIDT), Logistic Regression (LR) and a few types of Neuron Networks (NNs) [15], [16], [18].

2 Problem formulation

Clients' classification model is in Fig. 1. This model works with real PCs data from the period 2013-2015. The data matrix represents 1461 new clients – legal entities.

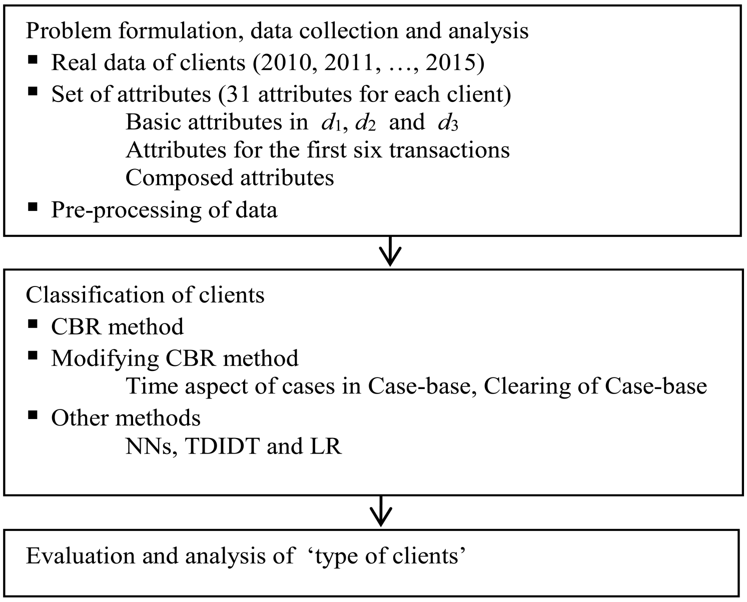
Problem formulation, data collection and analysis
▪ Real data of clients (2010, 2011, …, 2015)
▪ Set of attributes (31 attributes for each client)
Basic attributes in d_1, d_2 and d_3
Attributes for the first six transactions
Composed attributes
▪ Pre-processing of data
Classification of clients
▪ CBR method
▪ Modifying CBR method
Time aspect of cases in Case-base, Clearing of Case-base
▪ Other methods
NNs, TDIDT and LR
Evaluation and analysis of 'type of clients'

Fig. 1. The client classification model.

Each single client is described by the set of 31 attributes. These are:

– Evidence (1 attribute) – client's ID

– Time data (3 attributes) – Framework Agreement (FA) signature date d_1 (start of contractual relation), classification date d_2 (91 days from agreement signature), date of profitability evaluation d_3 (after meeting profitability conditions, no later than one year from FA signature)

– Data on the first six transactions of a client if they were realized prior to the date of classification, here each of the transactions includes: Transaction Volume, Transaction Profit and Time period from FA signature, from any previous transaction respectively (18 attributes)

- Composed attributes: Volume for the first 3, 6 and 12 months from FA signature (3 attributes), Profit for the first 3, 6 and 12 months from FA signature (3 attributes), Number of Transactions for the first 3, 6 and 12 months from FA signature (3 attributes).

Composed attributes for the first three months period (it means volume, profit and number of transactions) from FA signature date can be used in the framework of the classification. On the other hand these data serve primarily the purpose of calculation if a client is profitable or unprofitable like the other composed attributes.

The data was pre-processed. Normalization and standardization of data was a part of the pre-processing process as well as the division of data to training, testing and validation sets, and at the end elimination of 138 cases that did not realize any single transaction in period d_2 was done. These cases are in 94% cases unprofitable clients – classification of these clients is thus done under different conditions and such classification is not a part of regular company processes [12]. The training set thus included 686 cases, the testing set included 478 cases and the validation set included 159 cases.

Consequently the individual clients' classification models have been tested. The original classification system was used in the company since year 2011. The strong side of this system is simplicity (volume criteria on the first two transactions in period d_2), the negative side of this system is dramatic volume of errors, the system is error prone. In year 2015 this classification was, by means of statistical methods, re-designed. This re-design means that the average profit per d_2 is calculated based on the number of transactions and the average profit per transaction and thereby profit after one year from d_1 is forecasted. The quality of classification done in this way increased by 10%. Further more advanced classification methods have been tested. These methods are NNs, TDIDT and LR – Multi-layer perceptron (MLP), Radial Basis Function (RBF), C5, CRT, Quest and CHAID. Methods with the most exact results are presented in Table 1 [8], [9], [11].

Opposite to these methods a standard CBR model was created. In this model the case base was supplemented when the case had been evaluated as a new one. Thus we can talk here about a 'growing base of cases'. In time d_2 a client was classified according to the profitability of the nearest case in the cases base (here and after pattern). This classification was done based on 18 normalised and standardised attributes by means of calculation of space of standard Euclidean space [16]. This given case was consequently included into the base of cases. Until the time such case was allocated to profitable or to unprofitable, this happening in time d_3, it had not been used for the classification of new cases. CBR algorithm is, in this respect, more flexible than other classification tools because it is able to provide valid results already with a small number of cases in the base [13], p.10. While the above-stated methods (see Table 1) needed 686 cases in the training set, the CBR needed for the training set only a few tens of cases and it provided satisfactory results (about 75%) already on 300 cases in the case base (see Fig. 2). Consequently the quality of the classification stagnated, with the size of case base 900 cases it provided only 76.52% precision. It was only with 1000, 1100 respectively cases, that it provided solid classification precision values – 79.88%, 82.81% respectively [17].

Table 1. Results of clients' classification Source: adapted according to [12].

Classification method	Testing set Correct classification in %	Validation set Correct classification in %
Standard statistical methods	73.54	71.70
TDIDT C5 Classification	80.18	81.76
NN MLP Classification	82.22	83.65
LR Classification	79.25	78.62

The resulting values were however still relatively worse compared to the values acquired by means of TDIDT or NNs. When comparing values for validation sets the difference is even more visible. The CBR has the accuracy of mere 74.84%. The difference is thus five and more percentage points for the compared methods: MLP, C5 and LR. A problem was identified in case-basis where a part of deviated cases became invalid patterns.

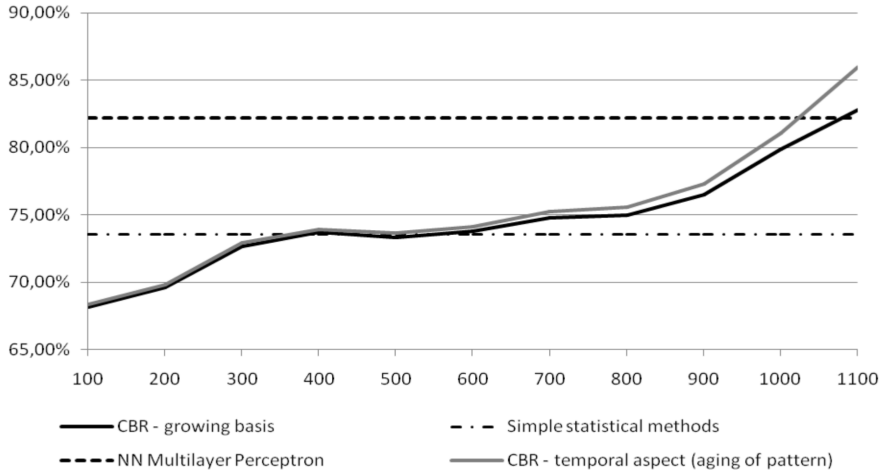

Fig. 2. Comparison of the precision of selected methods on the test set.

3 Problem solution

Two possible approaches were identified as a solution for classification quality. The first approach works with the assumption that patterns in the base age and they thus represent reality that was already overcome. Due to that incorrect classification of new cases can happen. The solution of this situation was possible only by an additional classification attribute (difference between i-th the new case d_{1NEW} and 'old' j-th case (pattern) d_{1CASE}. This approach proved to be ineffective. The quality of the prediction increased only by the order of tenths of a per cent (maximum was 1% with 1000 cases in the base) while this attribute required constant recalculation due to standardization of this attribute. Further it was possible to automatically eliminate the case when it reached a certain age limit. However, this approach did not prove to be useful for data covering a three years period be-cause it reduced the number of cases in the base and thereby also the quality of the classification that consequently did not exceed 75% [2], [5].

Another approach was chosen using constant in dependence on the difference between d_1 case and an example. Step by step space of time spaces was searched as well as space of the constant by means of iteration method Hill Climbing [15] where algorithm defined two local extremes 548 and 930 days. According to this method the time space was divided into three zones z_1, z_2 and z_3 by the following way:

$$z_1\text{: } (d_{1NEW} \text{ - } d_{1CASE}) < 548 \tag{1}$$
$$z_2\text{: } 548 \leq (\mathrm{d}_{1NEW} \text{ - } d_{1CASE}) \leq 930 \tag{2}$$
$$z_3\text{: } (d_{1NEW} \text{ - } d_{1CASE}) > 930 \tag{3}$$

The first zone constant is always 1. For the second zone local maximums were identified at values 1 and 1.4. For the third zone it was for value 3.6. By this approach higher quality results on the testing set were achieved (see Fig. 2, line CBR – temporal aspect). During validation there was decline to 77.99% for constants set at 1/1/3.6. The result is thus getting more and more accurate with growing base until it reaches difference 3.14% in favour of using the time aspect in CBR. In spite of good results on the validation set it is possible to see this use as experimental and its robustness must be further tested on solving other tasks.

Another possibility how to improve the quality of cases base was to assess the quality of cases and of patterns in time d_3. When they were identical, then they both entered the cases base. When they are not identical, then the following hap-pens:

- The similarity of cases is too small – both cases are valid and they enter the base

- The case entering the base is invalid – it is most likely a deviated case. This case does not enter the base

- The pattern is invalid – it is a deviated case that entered the base in the situation of small number of cases in case-base and based on a similarity with a very distant case, it shall be eliminated from the base

- The case and the pattern are invalid – the case does not enter the base and the pattern shall be eliminated from the base.

This approach can be also seen as a replacement for the time aspect because in case an pattern does not correspond to reality it is eliminated from the base. The advantage of this approach is that it completely eliminates deviated cases. The disadvantage is that the resolver views some cases as similar and they are not. Therefore their elimination and base reduction are not welcome. Further they are eliminated after d_3 of compared case. This time period can be up to one year long. During this period they can again become patterns while in the first approach they would be disadvantaged due to the time difference.

An important problem is to define the borderline where we can say that cases are so distant that their similarity may not lead to the same result. The more the given borderline shall grow the more cases shall be evaluated for being valid or not valid.

The algorithm itself inspected the environment of individual cases. When cases within the defined borderline led to the same results, as was the case of the researched case, then such case, in spite of its incorrect classification, was considered as valid. First the distance of the borderline was experimentally tested. The total number of tested cases was 396 (cases). When setting the distance to zero all cases remained in the base. As the borderline extended the number of cases eliminated from the base increased. Direct proportion does not quite apply here since in the same way as the distance in which we assess a case and an example grows then also the distance for which we assess their environment grows. At maximum 183 cases were eliminated from the base with values in interval from 2.495 to 2.554.

Table 2. Results of classification precision on the testing set in % by individual testing methods

Methods	Number of cases in the base								
	300	400	500	600	700	800	900	1000	1100
Standard CBR	72.69	73.69	73.34	73.76	74.78	75.00	76.52	79.88	82.81
Aging of cases	73.26	74.35	74.10	74.65	75.86	76.10	77.27	81.10	85.94
Aging of case and base clearing	74.89	75.38	75.96	76.70	78.21	78.40	79.41	84.02	86.27

The quality of classification increased with this action. However, it was essential to set an optimal borderline, to see if assessment of case aging definition changes respectively. Again the iteration method Hill Climbing was used. This time both on the testing set and on the validation set. The results of this method used on the validation set are presented in Fig. 3. The local maximum was always located on value 1.7 while in case of setting distance coefficients to 1/1/3.6 also on the interval from 2.353 to 2.571. Upon researching into spaces setting of coefficients to 1/1.4/3.6 was yet again evaluated as an optimum set. Results achieved on the test set can be compared in Table. 2. The result of validation set classification is 84.91% (when 169 examples are eliminated from the database) which is a better result than the result provided by NNs MLP (accuracy 83.65%)

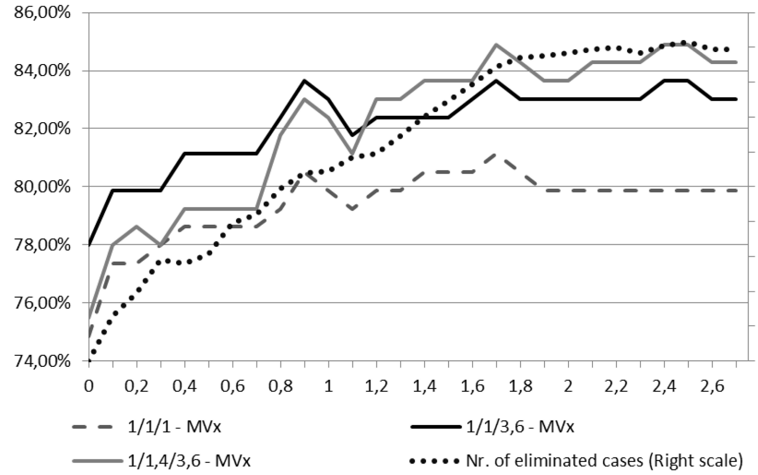

Fig. 3. The influence of distance setting on the classification's success (Validation Set) and the number of case eliminated from the base.

4 Conclusion

Based on the results the CBR model is a suitable possibility how to classify a client. The big advantage of this model is that this model starts to show good results already with a relatively small number of cases in the base (primarily in comparison with the original method of calculation). It can quickly show whether the generally used classification, used by a majority of trade companies, provides serious errors. Also as the number of cases in the base grows to substantial volume the CRB results are comparable with results of all other classification methods.

The results achieved by means of base clearing and by adding the time aspect to distance (similarity) of cases are very impressive. Both of the methods have their validity. Their potential importance should however be tested on further data. The same apply for Hill Climbing as a method for acquisition of parameters setting.

When comparing CBR with other methods the advantage of CBR is that it unambiguously refers to its example. So when it is essential to explain its decision then it is possible to show a concrete example. An expert at the same time may find out that the example is invalid for any possible reason, to eliminate it from the base and to start the resolver one more time.

Compared to TDIDT CBR has (in the same way as NNs) the disadvantage that it does not create a network of rules. Such rules, with a certain level of simplification, could be used for instance for a proposal of rewards for sales representatives for clients acquisition.

Further research into this area should be focused on the area of the robustness of the model. Either it is possible to improve model adaptation features toward data or to make the phase of search for a suitable pattern in base more accurate. Here it is possible to use weighting of attributes based on their importance (it is defined either by an expert or by an adaptive algorithm) or to use any method working with ambiguity such as are, for instance, fuzzy or rough sets [10], [13].

Acknowledgments. This article was supported by the project No. SGS–2017–019 with title Models Synthesis and Analysis for Implementation Support of Smart Cities and Regions Concept at the Faculty of Economics and Administration, University of Pardubice.

References

1. Aamodt, A., Plaza, E.: Case-based reasonning: foundational issues, methodological variations and system aproach. AI Communications. 7 (1), 39–59 (1994)
2. Ahn, H., Kim, K.J.: Global optimization of case-based reasoning for breast cytology diag-nosis. Expert Systems with Applications. 36 (1), 724–734 (2009)
3. Besbes, G., Baazaoui-Zghal, H.: Modular ontologies and CBR-based hybrid system for web information retrieval. Multimed Tools Appl. 74 (18), 8053–8077 (2015)
4. Cho, B. et al.: CBR-based network performance management with multi-agent approach. Cluster Comput. 20 (1), 757–767 (2017)
5. Cunningham, P., Zenobi, G.: Case representation issues for case-based reasoning from ensemble research. In: International Conference on Case-Based Reasoning. Springer Berlin, Heidelberg, pp. 146–157 (2001)
6. Finnie, G.; Sun, Z.: R5 model for case-based reasoning. Knowledge-Based Systems. 16 (1), 59–65 (2003)
7. Guo, Y., Hu, J., Peng, Y.: Research of new strategies for improving CBR sys-tem. Artif Intell Rev. 42 (1), 1–20 (2014)
8. Kasparova, M., Krupka J.: Air Quality Modelling by Decision Trees in the Czech Republic Locality. In: The 8th WSEAS Int. Conf. on Applied Informatics and Communications (AIC'08), pp. 196–201. WSEAS Press, Greece (2008)
9. Kvasnicka, V. et.al. Uvod do teorie neuronovych sieti. Bratislava, IRIS (1997)
10. Mezera, F., Krupka, J.: Local model of the air quality on the basis of rough sets theory. In: Snasel V., Abraham A., Corchado E. (eds.) Soft Computing Models in Industrial and Environmental Applications. AISC, vol. 188, pp. 277–286. Springer, Berlin, Heidelberg (2013)

11. Mezera, F., Krupka, J.: Environmental Modelling Based on Rough-Fuzzy Approach. In: Gruca, A., Czachorski, T., Kozielski, S. (eds.) Man-Machine Interactions 3. AISC, vol. 242, pp. 407-414. Springer, Cham (2014)
12. Mezera, F., Krupka, J.: Decision-Making Support and Its Application in Public Administration. In: Proc. of the 11th Int. Scientific Conf. Public Administration 2016, 22. 9. 2016, Pardubice: University of Pardubice, pp. 181–188 (2016)
13. Pal, S.K., Shiu S.C.K.: Foundation of Soft Case-Based Reasoning. Hobokren, New Jersey (2004)
14. Perner P.: Case-Based Reasoning and the Statistical Challenges. In: Althoff K. et al. (eds.) Advances in Case-Based Reasoning. ECCBR 2008. LNCS, vol. 5239, pp. 430-443. Springer, Berlin, Heidelberg (2008)
15. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach (2nd ed.), Prentice Hall, Upper Saddle River, New Jersey, pp. 111–114 (2003)
16. Soltes, E.: Regresna a korelacna analyza s aplikaciami. Jura Edition, Bratislava (2008)
17. Sun, Z., Han, J., Dong, D.: Five Perspectives on Case-based reasoning. In: Huang, D.S. et al. (eds.) Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. LNAI, vol. 5227, pp. 410–419. Springer, Berlin, Heidelberg (2008)
18. Watson, I.: Applying Case-Based Reasoning: Techniques for Enterprise Systems. Morgan Kaufmann Publisher, Inc., San Francisco (1997)
19. Xu, M., Yu, H., Shen, J.: New algorithm for CBR-RBR fusion with robust thresholds. Chinese Journal of Mechanical Engineering. 25 (6), 1255–1265 (2012)
20. Zhu, Z. et al.: Literature review on the creativity of CBR applications. Artif In-tell Rev. 40 (4), 379–390 (2013)