Dominance-based Rough Set Approach to Bank Customer Satisfaction Analysis

Marcin Szeląg[1] and Roman Słowiński[1,2]

[1] Institute of Computing Science, Poznań University of Technology, Poznań, Poland
[2] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
{mszelag,rslowinski}@cs.put.poznan.pl

Abstract. We present an application of the Dominance-based Rough Set Approach (DRSA) to analysis of bank data concerning customer satisfaction. The analysis is conducted using two new applications – RuLeStudio and RuleVisualization. The first one is designed to experiment with different rule models, depending on chosen settings. The second one is used for visualization and in-depth examination of the constructed model. Our analysis gives insight into the data in terms of monotonic decision rules which describe loyal customers and the ones who ceased cooperation with the bank. Such analysis is in vain of explainable AI, aiming to obtain a transparent decision model, that can be understood by decision makers. We also compare predictive performance of our approach with some well-known machine learning methods.

Keywords: Dominance-based Rough Set Approach · Ordinal classification · Monotonic relationships · Decision rules · Customer Satisfaction.

1 Introduction

Rough set theory (RST) was introduced by Zdzisław Pawlak [5]. Since then, different extensions and applications have been proposed. An important direction of research, initiated by Greco, Matarazzo, and Słowiński, concerns adaptation of RST to multicriteria decision aiding. They proposed the Dominance-based Rough Set Approach (DRSA) [4], which employs dominance instead of indiscernibility relation among objects in the definition of rough approximations, and builds decision models in terms of monotonic $if \ldots, then \ldots$ decision rules. DRSA is able to take into account monotonic relationships present in data between condition and decision attributes. Rule models are considered to be both transparent for a user, and useful for explanation of suggested decisions, which is an important aspect of AI methods, apart from sole predictive performance.

In this study, we show an application of DRSA to analysis of a bank customer data. Employing decision rules, we wish to present readable patterns of customers who left the bank. We use Variable Consistency DRSA (VC-DRSA) [2] and introduce a new rule classifier. We also use a new software designed to learn, explore and apply decision rules.

In Section 2, we describe the methodology. Section 3 presents the analysis of public domain data obtained from a bank. The last Section 4 groups conclusions.

2 Methodological Background

We employ VC-DRSA with cost-type consistency measure ϵ [2], used to calculate lower approximations of unions of ordered decision classes, and to induce decision rules from these approximations. This involves threshold $\theta_X \in [0, 1]$ to be defined by the user for each upward/downward union $X \subseteq U$, where U is a universe of (learning) objects. Then, the lower approximation of X is composed of objects whose consistency (measured by ϵ) is not worse than θ_X. The rules are induced using *VC-DomLEM* algorithm [3], and later, those with confidence ≤ 0.5 are removed. They explain observed decisions, and can classify any new object.

In DRSA, classification of an object based on matching rules can be done in different ways (see, e.g., [1]). We propose a *mode classifier* being able to resolve conflicting class assignments. It is implemented in RuLeStudio[3]. Let consider the set of objects described in terms of two gain-type criteria g_1, g_2 shown in Fig. 1. Let denote class i by Cl_i. Object z to be classified (red cross) is covered by rules r_1, r_2, r_3, r_4, suggesting, respectively: "at least Cl_5", "at least Cl_4", "at most Cl_1", and "at most Cl_2". Then: (i) *upward intersection* is "at least Cl_5", (ii) the most prudent upward class is Cl_5, (iii) *downward intersection* is "at most Cl_1", (iv) the most prudent downward class is Cl_1, (v) *mode* of the two classes is computed. Observe that r_1 covers 2 objects from Cl_5, and r_2 covers 1 additional object from Cl_5. Then, Cl_5 is supported by 3 objects. Moreover, r_3 covers 2 objects from Cl_1, and r_4 covers no additional object from Cl_1. Then, 2 objects support Cl_1. Consequently, Cl_5 is returned by the classifier (more frequent class).

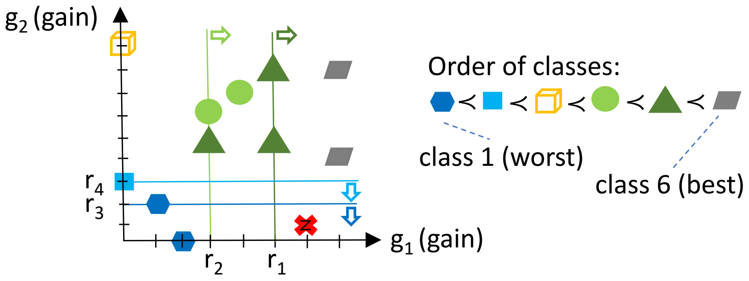


Fig. 1. Exemplary set of objects for illustration of the *mode classifier*

If no rule matches z, one can suggest a majority class (optimizing classification accuracy) or median class (optimizing mean absolute error).

3 Case Study of Bank Customer Satisfaction

We analyze the *churn* data set publicly available at kaggle.com[4], featuring 10 condition attributes, incl. 4 continuous ones. To build a balanced universe of

[3] www.cs.put.poznan.pl/mszelag/Software/RuLeStudio/RuLeStudio.html
[4] https://www.kaggle.com/mathchi/churn-for-bank-customers

objects U, we drew 2000 churning customers ($Exited = 1$) and 2000 loyal customers ($Exited = 0$)[5]. In this study, we compare our method (ϵ-VC-DRSA + mode classifier; $Exited = 0$ as default decision) with three ML classifiers available in WEKA[6] (with default parameters): SVM (SMO) with polynomial kernel, C4.5 (J48) tree classifier, and naive Bayes (NaiveBayes) classifier. We estimate predictive performance using classification accuracy.

We considered the remarks at kaggle.com, WEKA's histograms, and trial-and-error assessment in RuLeStudio to assign attribute preference orders as follows: *CreditScore* – gain (after kaggle.com), *Geography* – none (nominal attribute), *Gender* – none (nominal attribute), *Age* – cost (distribution for class $Exited = 1$ shifted to the right), *Tenure* – cost (verified in RuLeStudio), *Balance* – gain (kaggle.com), *NumOfProducts* – we duplicated this attribute and assigned type gain to the first clone, and type cost to the second one (the histogram shows prevalence of loyal customers when $NumOfProducts = 2$, and the opposite otherwise), *HasCrCard* – none (nominal attribute), *IsActiveMember* – gain (kaggle.com) *EstimatedSalary* – gain (kaggle.com). For the decision attribute *Exited*, label 0 was more preferred than 1 (bank's viewpoint).

In our study, unions of classes boil down to single classes – characterized by decisions $Exited = 0$ and $Exited = 1$. We assumed a common threshold θ_X for both classes. Using cross-validation in RuLeStudio, we tested thresholds 0, 0.01, 0.02, and 0.05, choosing value 0.01. Note that for $\theta_X = 0$ (classical DRSA), the quality of classification was 0.68775, while for $\theta_X = 0.01$ it increased to 0.996.

Table 1 presents comparison of average classification accuracy from 3 independent runs of 10-fold cross-validation. One can see that our method performed better than SVM, slightly better than naive Bayes, and slightly worse than C4.5. Next, we analyzed the models trained on all 4000 objects. Reclassification accuracy was: SVM 70.225%, naive Bayes 72.25%, C4.5 85.525%, our approach 83.825% (2nd best). C4.5 tree size was equal to 320 with 164 leaves. The tree had many long paths which were hard to understand and did not respect the above preference orders. When transformed to 164 rules, even after aggregating redundant conditions (e.g., Age ≤ 41 and Age ≤ 37 resulted in Age ≤ 37), average rule length was 7.81 and average rule support was 24.39. The model learned by ϵ-VC-DRSA contained 770 rules. We explored them in RuleVisualization[7]. Our observations: (i) on avg. 5.91 conditions per rule – much better than C4.5; (ii) avg. rule support 34.1 – again much better than C4.5; (iii) top

Table 1. Comparison of average classification accuracy in 3 × 10-fold cross-validation

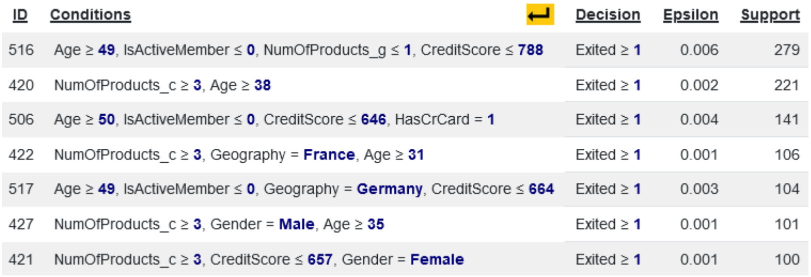
method	**ϵ-VC-DRSA+mode**	SVM	C4.5	naive Bayes
avg. accuracy	73.25	69.91	75.18	71.87
rank	2	4	1	3

[5] http://www.cs.put.poznan.pl/mszelag/Research/bank-churn

[6] https://www.cs.waikato.ac.nz/~ml/weka; used version: 3.8.6

[7] www.cs.put.poznan.pl/mszelag/Software/RuleVisualization/RuleVisualization.html

attributes present in rules are: Geography (in 76.2% of rules), Age (in 74.9% of rules), EstimatedSalary (in 59.9% of rules), CreditScore (in 58.7% of rules); (iv) most often co-occurence of attributes concerns Geography and Age; (v) the two strongest rules concern decision Exited ≥ 1 and are supported by 279 and 221 objects. Fig. 2 shows top rules for customers who left the bank (Support ≥ 100, Confidence ≥ 0.95). Remark that NumOfProducts ≥ 3 is often related to churn.



ID	Conditions ↵	Decision	Epsilon	Support
516	Age ≥ **49**, IsActiveMember ≤ **0**, NumOfProducts_g ≤ **1**, CreditScore ≤ **788**	Exited ≥ **1**	0.006	279
420	NumOfProducts_c ≥ **3**, Age ≥ **38**	Exited ≥ **1**	0.002	221
506	Age ≥ **50**, IsActiveMember ≤ **0**, CreditScore ≤ **646**, HasCrCard = **1**	Exited ≥ **1**	0.004	141
422	NumOfProducts_c ≥ **3**, Geography = **France**, Age ≥ **31**	Exited ≥ **1**	0.001	106
517	Age ≥ **49**, IsActiveMember ≤ **0**, Geography = **Germany**, CreditScore ≤ **664**	Exited ≥ **1**	0.003	104
427	NumOfProducts_c ≥ **3**, Gender = **Male**, Age ≥ **35**	Exited ≥ **1**	0.001	101
421	NumOfProducts_c ≥ **3**, CreditScore ≤ **657**, Gender = **Female**	Exited ≥ **1**	0.001	100

Fig. 2. Top rules describing customers who ended cooperation with the bank

4 Conclusions

In this paper, we analysed customer satisfaction data from a bank using Variable Consistency Dominance-based Rough Set Approach, and some reference machine learning methods. We employed two new programs suitable for this task: RuLeStudio and RuleVisualization. Moreover, we proposed a new rule classification strategy (mode classifier), implemented in RuLeStudio. The results obtained using our approach are competitive with respect to average classification accuracy, but even more important, the induced rule model gives a clear insight into the problem, helping the bank to improve long-term customer relationships.

Acknowledgements This research was supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

1. Błaszczyński, J., Greco, S., Słowiński, R.: Multi-criteria classification – a new scheme for application of dominance-based decision rules. European Journal of Operational Research **181**(3), 1030–1044 (2007)
2. Błaszczyński, J., Greco, S., Słowiński, R., Szeląg, M.: Monotonic variable consistency rough set approaches. Int. J. of Approx. Reasoning **50**(7), 979–999 (2009)
3. Błaszczyński, J., Słowiński, R., Szeląg, M.: Sequential covering rule induction algorithm for variable consistency rough set approaches. Information Sciences **181**, 987–1002 (2011). https://doi.org/10.1016/j.ins.2010.10.030
4. Greco, S., Matarazzo, B., Słowiński, R.: Rough sets theory for multicriteria decision analysis. European Journal of Operational Research **129**(1), 1–47 (2001)
5. Pawlak, Z.: Rough sets. Int. J. of Inf. & Computer Sciences **11**, 341–356 (1982)