Dominance-based Rough Set Approach to Bank Customer Satisfaction Analysis

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

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

PP-RAI'2022, Gdynia, May 26

Outline

Introduction

Problem description

- We analyzed the **churn** data set publicly available at kaggle.com[1]; 10 condition attributes, incl. 4 continuous ones.
- This is a **binary** problem; $Exited = 1$ denotes a customer who leaved the bank, $Exited = 0$ denotes a loyal one.
- Original class distribution: $Exited = 1$: 2037 objects, $Exited = 0$: 7963 objects.
- **Balanced** subproblem used in this study: $Exited = 1$: 2000 objects, $Exited = 0$: 2000 objects (random selection).
- Data used in the case study: http://www.cs.put.poznan.pl/mszelag/Research/bank-churn (can be used to reproduce experiments).
- Predictive performance estimated using **classification accuracy**.

[1] https://www.kaggle.com/mathchi/churn-for-bank-customers

Sample of the original data

No.	CreditScore (condition,active	Geography (condition,active)	Gender (condition,active)	Age (condition,active)	Tenure (condition,active)	Balance (condition,active)	NumOfProducts (condition,active)	HasCrCard (condition,active)	IsActiveMember (condition,active)	EstimatedSalary (condition,active)	Exited (decision,active)
1	619	France	Female	42	2	0.0	1	1	1	101348.88	1
2	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	699	France	Female	39	1	0.0	2	0	0	93826.63	0
5	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
6	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	822	France	Male	50	7	0.0	2	1	1	10062.8	0
8	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	501	France	Male	44	4	142051.07	2	0	1	74940.5	0
10	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
11	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
12	497	Spain	Male	24	3	0.0	2	1	0	76390.01	0
13	476	France	Female	34	10	0.0	2	1	0	26260.98	0
14	549	France	Female	25	5	0.0	2	0	0	190857.79	0
15	635	Spain	Female	35	7	0.0	2	1	1	65951.65	0
16	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
17	653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1
18	549	Spain	Female	24	9	0.0	2	1	1	14406.41	0

At kaggle.com one can find some **hints** about **preference orders**.

Motivation for the case study

- From the view point of the bank, it is much more expensive to sign in a new client than keeping an existing one.
- It is advantageous for banks to know what leads a client towards the decision to leave the bank.
- Churn prevention allows companies to develop loyalty programs and retention campaigns to keep customers.
- As there are some **preference orders** and **inconsistencies w.r.t. dominance** involved, we decided to apply **Variable Consistency Dominance-based Rough Set Approach** (VC-DRSA) to develop **explainable decision rule model**.
- We compared performance of VC-DRSA with **three competing ML classifiers** available in WEKA (with default parameters): **SVM** (SMO) with polynomial kernel, **C4.5** (J48) tree classifier, and **naive Bayes** (NaiveBayes) classifier.

Methodological background

Ordinal classification with monotonicity constraints

	buying	maint	doors	persons	lug_boot	safety	class	
	vhigh	vhigh	2	2	small	med	unacc	X_1
	med	vhigh	3	more	small	med	unacc	
	vhigh	high	2	4	med	low	unacc	
	vhigh	high	2	4	big	low	unacc	
	med	low	2	4	big	low	unacc	
y	low	low	4	more	big	high	unacc	
	high	med	2	more	med	high	acc	X_2
	med	vhigh	3	more	med	med	acc	
	med	vhigh	3	more	med	high	acc	
	med	vhigh	3	more	big	med	acc	
	med	vhigh	3	more	big	high	acc	
	low	low	4	more	small	med	acc	
x	low	low	2	more	big	med	good	X_3
	low	low	4	more	small	high	good	
	low	low	4	more	big	med	good	
	med	med	4	more	med	high	vgood	X_4
	med	low	2	4	big	high	vgood	
	low	low	4	more	big	high	vgood	
	q_1	q_2	q_3	q_4	q_5	q_6	dec	

$$\forall q \in C, y \succeq_q x \Leftrightarrow \begin{cases} yDx \\ x\,d\,y \end{cases} \Leftrightarrow \begin{cases} y \in D^+(x) \\ x \in D^-(y) \end{cases} \qquad dec(y) \prec dec(x)$$

Dominance-based Rough Set Approach (DRSA)

In **Classical DRSA** (CDRSA)[2], **lower approximations** of **unions of ordered classes** $X_i^{\geq}$, $X_i^{\leq}$ are defined using strict inclusion relation:

$$\underline{X_i^{\geq}} = \{x \in U : D^{+}(x) \subseteq X_i^{\geq}\}, \tag{1}$$

$$\underline{X_i^{\leq}} = \{x \in U : D^{-}(x) \subseteq X_i^{\leq}\}. \tag{2}$$

Upper approximations of $X_i^{\geq}$, $X_i^{\leq}$ are defined as

$$\overline{X_i^{\geq}} = \{x \in U : D^{-}(x) \cap X_i^{\geq} \neq \emptyset\}, \tag{3}$$

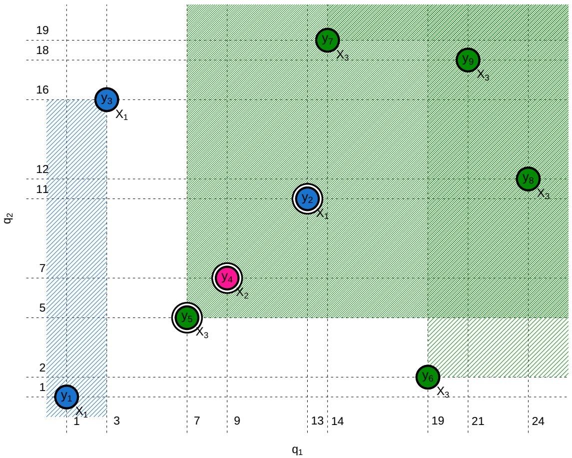
$$\overline{X_i^{\leq}} = \{x \in U : D^{+}(x) \cap X_i^{\leq} \neq \emptyset\}. \tag{4}$$

[2] S. Greco, B. Matarazzo, R. Słowiński, Rough Sets Theory for Multicriteria Decision Analysis. European Journal of Operational Research, 129(1), 2001, pp. 1-47

In **Variable Consistency DRSA** (VC-DRSA), lower approximations are defined using **object consistency measures**.

E.g., often used **cost-type** consistency measures $\epsilon_{X_i^{\geq}} : U \rightarrow [0,1]$, $\epsilon_{X_i^{\leq}} : U \rightarrow [0,1]$ are defined as:

$$\epsilon_{X_i^{\geq}}(x) = \frac{|D^{+}(x) \cap \neg X_i^{\geq}|}{|\neg X_i^{\geq}|}, \quad \epsilon_{X_i^{\leq}}(x) = \frac{|D^{-}(x) \cap \neg X_i^{\leq}|}{|\neg X_i^{\leq}|}. \quad (5)$$



$$\epsilon_{X_3^{\geq}}(y_5) = \frac{2}{4}, \qquad \epsilon_{X_1^{\leq}}(y_3) = 0$$

Dominance-based Rough Set Approach (DRSA)

Applying measure ϵ, **probabilistic lower approximations** of $X_i^{\geq}$, $X_i^{\leq}$ are defined as

$$\underline{X_i^{\geq}} = \{x \in X_i^{\geq} : \epsilon_{X_i^{\geq}}(x) \leq \theta_{X_i^{\geq}}\}, \tag{6}$$

$$\underline{X_i^{\leq}} = \{x \in X_i^{\leq} : \epsilon_{X_i^{\leq}}(x) \leq \theta_{X_i^{\leq}}\}, \tag{7}$$

where **thresholds** $\theta_{X_i^{\geq}}, \theta_{X_i^{\leq}} \in [0, 1)$.

The above definitions constitute approach called ϵ-**VC-DRSA**.

Dominance-based Rough Set Approach (DRSA)

ϵ-**VC-DRSA** offers good properties[3], as measure ϵ is both **monotonic w.r.t. set of attributes** (m1) and **monotonic w.r.t. dominance** (m4), which is not the case, e.g., for rough membership object consistency measure μ.

Advantage of (VC-)DRSA - **no need for discretization** in case of numerical attributes!

[3] J. Błaszczyński, S. Greco, R. Słowiński, M. Szeląg, Monotonic Variable Consistency Rough Set Approaches. International Journal of Approximate Reasoning, 50(7), 2009, pp. 979-999

Transformation of data with (some) regular attributes

Possible actions for a **regular attribute** $q \in C$:

- "leave attribute as-is": if $q(y) = q(x)$, then $y \succeq_q x$, otherwise relation does not hold,
- "process attribute"[4]:
  - "**duplicate+impose**" (only for numerical attributes) $\Rightarrow$ original attribute replaced with 2 criteria (one gain criterion and one cost criterion),
  - "**binarize**" (only for nominal attributes with 3+ domain values) $\Rightarrow$ original attribute with v different values replaced with v binary (0/1) regular attributes.

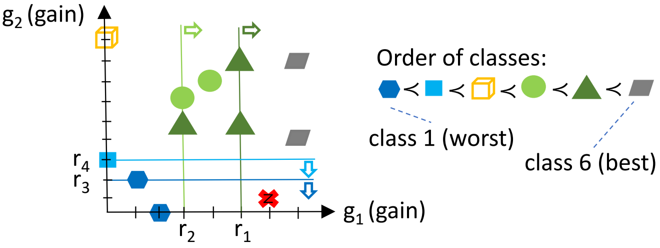
[4] J. Błaszczyński, S. Greco, R. Słowiński. Inductive discovery of laws using monotonic rules. Engineering Applications of Artificial Intelligence, 25:284–294, 2012.

Decision rule induction

- Rules are induced using **VC-DomLEM** sequential covering algorithm[5].
- When using consistency measure ϵ, rule induction is **fast** due to exploitation of two properties:
  - **(m1)** (when inducing a rule, each attribute is tested once),
  - **(m4)** (when inducing a rule, not all conditions on the current attribute need to be checked – shrinking window technique).

Rules with confidence ≤ 0.5 are removed (avoids overfitting).

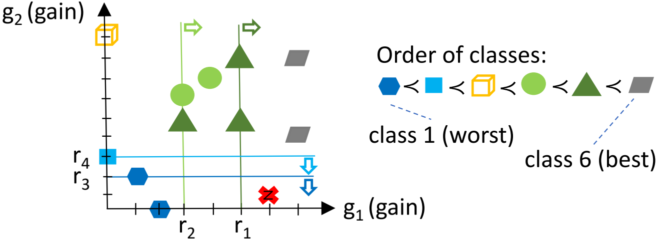
[5] J. Błaszczyński, R. Słowiński, M. Szeląg, Sequential Covering Rule Induction Algorithm for Variable Consistency Rough Set Approaches. Information Sciences, 181, 2011, pp. 987-1002



Object z to be classified is **covered** by rules: r_1 (decision "at least X_5"), r_2 (decision "at least X_4"), r_3 (decision "at most X_1"), and r_4 (decision "at most X_2").

Then: (i) **upward intersection** is "at least X_5", (ii) the most prudent upward class is X_5, (iii) **downward intersection** is "at most X_1", (iv) the most prudent downward class is X_1, (v) **mode** of the two classes is computed.

New rule classification strategy - mode classifier



Observe that r_1 covers 2 objects from X_5, and r_2 covers 1 additional object from X_5. Then, X_5 **is supported by 3 objects**. Moreover, r_3 covers 2 objects from X_1, and r_4 covers no additional object from X_1. Then, class X_1 **is supported by 2 objects**.

Consequently, X_5 is returned by the classifier (more frequent class).

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

Case study of bank customer satisfaction

Software involed in the case study

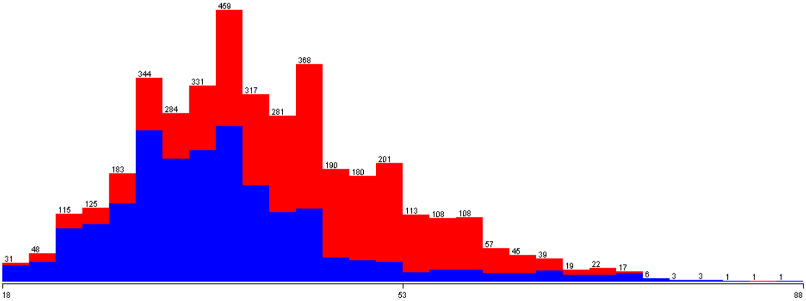
- In ϵ-VC-DRSA, we took $Exited = 0$ as **default decision** (when no rule covers test object).
- We used two new applications supporting (VC-)DRSA: **RuLeStudio**[6] and **RuleVisualization**[7].
- **RuLeStudio** (replacement for jMAF) – data consistency checking, rule induction and application (also using mode classifier), basic inspection of rules, cross-validation; handles analysis of data with missing attribute values.
- **RuleVisualization** – exploration and visualization of induced decision rules.
- Both programs are based on open-source **ruleLearn** library[8].
- Competitive methods were run in **WEKA** (version 3.8.6).

---

[6] www.cs.put.poznan.pl/mszelag/Software/RuLeStudio/RuLeStudio.html
[7] www.cs.put.poznan.pl/mszelag/Software/RuleVisualization/RuleVisualization.html
[8] github.com/ruleLearn/rulelearn

Assessment of attribute preference orders

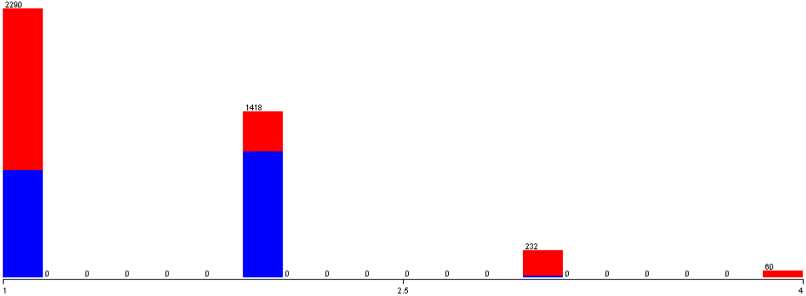
- We considered the remarks at kaggle.com, WEKA's histograms, and trial-and-error assessment in RuLeStudio to **assign attribute preference orders** as follows:
  - **CreditScore** – gain (after kaggle.com),
  - **Geography** – none (nominal attribute),
  - **Gender** – none (nominal attribute),
  - **Age** – cost (distribution for $Exited = 1$ shifted to the right),



  - **Tenure** – cost (in RuLeStudio gave better acc. in 10-fold CV),

Assessment of attribute preference orders

- We considered the remarks at kaggle.com, WEKA's histograms, and trial-and-error assessment in RuLeStudio to **assign attribute preference orders** as follows:
  - **Balance** – gain (after 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),



Assessment of attribute preference orders

- We considered the remarks at kaggle.com, WEKA's histograms, and trial-and-error assessment in RuLeStudio to **assign attribute preference orders** as follows:
  - **HasCrCard** – none (nominal attribute),
  - **IsActiveMember** – gain (after kaggle.com),
  - **EstimatedSalary** – gain (after kaggle.com).
- For the decision attribute **Exited**, label 0 was more preferred than 1 (bank's viewpoint).

Sample of the case study data

No.	CreditScore (condition,active)	Geography (condition,active)	Gender (condition,active)	Age (condition,active)	Tenure (condition,active)	Balance (condition,active)	NumOfProducts_g (condition,active)	NumOfProducts_c (condition,active)	HasCrCard (condition,active)	IsActiveMember (condition,active)	EstimatedSalary (condition,active)	Exited (decision,active)
1	584	Germany	Male	42	3	137479.13	1	1	1	0	25669.1	0
2	660	Germany	Male	39	9	134599.33	2	2	1	0	183095.87	0
3	676	Spain	Female	30	5	0.0	2	2	0	1	157888.5	0
4	561	Spain	Male	28	6	123692.0	1	1	1	1	70548.96	0
5	696	France	Female	30	8	0.0	2	2	1	1	196134.44	0
6	757	Germany	Male	33	1	122088.67	1	1	1	0	42581.09	0
7	545	France	Male	30	3	0.0	2	2	1	0	170307.43	0
8	723	France	Male	42	2	99095.73	1	1	1	1	17512.53	0
9	650	France	Female	43	6	0.0	2	2	1	1	16301.91	0
10	717	France	Male	28	4	128206.79	1	1	1	1	54272.12	0
11	521	France	Female	32	2	136555.01	2	2	1	1	129353.21	0
12	651	France	Male	28	7	0.0	2	2	1	0	823.96	0
13	700	France	Female	30	9	0.0	1	1	1	1	174971.64	0
14	675	Spain	Male	33	3	0.0	2	2	1	0	45348.08	0
15	628	France	Male	34	4	158741.43	2	2	1	1	126192.54	0
16	643	Spain	Female	35	6	0.0	2	2	1	1	41549.64	0
17	779	Spain	Male	33	3	0.0	2	2	1	0	30804.68	0
18	710	Spain	Female	38	4	0.0	2	2	1	1	136390.88	0

Comparing to kaggle.com, **we changed preference orders for four attributes**.

Assessment of ϵ-VC-DRSA consistency threshold

- For **binary classification**, 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** (gives the best avg. accuracy).
- 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**.

Comparison of avg. classification accuracy in 3×10-fold CV

Tablica: 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

- It is possible **improve naive Bayes** by enabling discretization for numeric attributes (-D switch).
  Then, avg. classification accuracy increases to **75.96%**.
- Other two competing classifiers have **too many parameters** to be tuned manually.

Comparison of models trained on all 4000 objects

Tablica: Comparison of classification accuracy in **reclassification** [%]

Method	ϵ-**VC-DRSA+mode**	SVM	C4.5	Naive Bayes
Accuracy	83.825	70.225	85.525	72.25

- It is possible **improve naive Bayes** by enabling discretization for numeric attributes (-D switch).
  Then, classification accuracy increases to **76.525%**.
- Other two competing classifiers have **too many parameters** to be tuned manually.

Comparison of ϵ-VC-DRSA rules and C4.5 tree

C4.5 tree:

- size was equal to 320 with 164 leaves,
- many long paths which were **hard to understand**,
- **did not respect** the above preference orders,
- when transformed to **164 rules**, average rule length was **7.81** and average rule support was **24.39**.

ϵ-VC-DRSA rules:

- **770 rules** (after removing rules with conf. $\leq$ 0.5),
- avg. rule length was **5.91** – much better than C4.5,
- avg. rule support was **34.1** – again much better than C4.5,
- **top attributes** in rules: Geography (in 76.2% of rules), Age (74.9%), EstimatedSalary (59.9%), CreditScore (58.7%),
- most often **co-occurence** of attributes: Geography and Age,
- support of 2 **strongest rules** (Exited ≥ 1): 279 & 221 obj.

Top rules for customers who left the bank

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

Rysunek: Top rules describing customers who ended cooperation with the bank (support $\geq$ 100, confidence $\geq$ 0.95)

NumOfProducts ≥ 3 is often related to churn.

Conclusions

Conclusions

- We analysed customer satisfaction data from a bank using **VC-DRSA**, and three **reference ML methods**.
- We employed two new programs suitable for this task: **RuLeStudio** and **RuleVisualization**.
- 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.
- The induced **rule model** gives a **clear insight into the problem**, helping the bank to improve long-term customer relationships.

Questions and discussion

Acknowledgements

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