sciendo

Number of Instances for Reliable Feature Ranking in a Given Problem

Marko Bohanec
Salvirt Ltd., Ljubljana, Slovenia
Mirjana Kljajić Borštnar
Faculty of Organizational Sciences, University of Maribor, Kranj, Slovenia
Marko Robnik-Šikonja
Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Abstract

Background: In practical use of machine learning models, users may add new features to an existing classification model, reflecting their (changed) empirical understanding of a field. New features potentially increase classification accuracy of the model or improve its interpretability. **Objectives:** We have introduced a guideline for determination of the sample size needed to reliably estimate the impact of a new feature. **Methods/Approach:** Our approach is based on the feature evaluation measure ReliefF and the bootstrap-based estimation of confidence intervals for feature ranks. **Results:** We test our approach using real world qualitative business-to-business sales forecasting data and two UCI data sets, one with missing values. The results show that new features with a high or a low rank can be detected using a relatively small number of instances, but features ranked near the border of useful features need larger samples to determine their impact. **Conclusions:** A combination of the feature evaluation measure ReliefF and the bootstrap-based estimation of confidence intervals can be used to reliably estimate the impact of a new feature in a given problem.

Keywords: machine learning, feature ranking, feature evaluation
JEL classification: C61
Paper type: Research article

Received: Jan 31, 2018
Accepted: Apr 21, 2018

Citation: Bohanec, M., Kljajić Borštnar, M., Robnik-Šikonja, M. (2018), "Number of Instances for Reliable Feature Ranking in a Given Problem", Business Systems Research, Vol. 9, No. 2, pp. 35-44.
DOI: 10.2478/bsrj-2018-0017

Acknowledgments: We are grateful to the company Salvirt Ltd. for funding the research and development of the optimization algorithm, used in this paper. Mirjana Kljajić Borštnar and Marko Robnik-Šikonja were supported by the Slovenian Research Agency, ARRS, through research programmes P5-0018 and P2-0209, respectively.

Introduction

In business practice, users of machine learning (ML) models are pragmatic about their effort to collect data describing a business process, for example selling into business-to-business (B2B) market segment. As mentioned in (Bohanec et al., 2016), users are upfront interested to learn how many historic cases are needed for the model to identify the most relevant features. For example, in Bohanec et al. (2016) only ≈1/3 of the final data set would be needed to identify top three features with 80% certainty (if their rank within top 3 is not relevant).

When the data set is collected and the model built, optimized and in use, a new question arises from domain-expert users, adding new features *ad hoc* (Guyon et al., 2003): how many instances are needed to estimate the impact of a new, candidate feature? Here users try to minimize the effort needed, which in practice means that only a few dozen of instances could be available for an assessment of feature's impact. In this paper, we extend our previous research to answer this question.

In (Bohanec et al., 2016) we analyzed the number of features and the number of instances needed to learn important features in a general business setting. Here we focus on reliability of ranks for new features given the context of an existing data set. We use a publicly available B2B sales forecasting data set (Bohanec, 2017) as a case study. We report the summary of applying the presented approach to two additional data sets, Wine (Forina et al., 1991) and Chronic Kidney Disease (Soundarapandian, 2015) (CKD), available from data repository at University of California, Irvine, US (UCI). To reliably estimate the impact of a new feature in a given problem, described with a data set, we combine feature evaluation measure ReliefF and bootstrap-sampled confidence intervals.

In contrast to this work, the majority of previous studies on sample size focused on the relationship between sample size and model performance. For example, Beleites et al. (2013) established that the sample size is related to the learning curve of classifier's model performance in Raman spectroscopic five class classification problem. The relationship between sample size and model's performance for B2B sales prediction problem was visually indicated in (Bohanec et al., 2015b). Figueroa et al. (2012) propose a sample size prediction algorithm that conducts weighted fitting of learning curves on clinical text and waveform classification tasks.

The rest of the paper is organized as follows. In Section 2 we introduce B2B data set and calculate ground truth. In Section 3 we formalize the problem, and continue with experiments in Section 4. Our conclusions are put forward in Section 5.

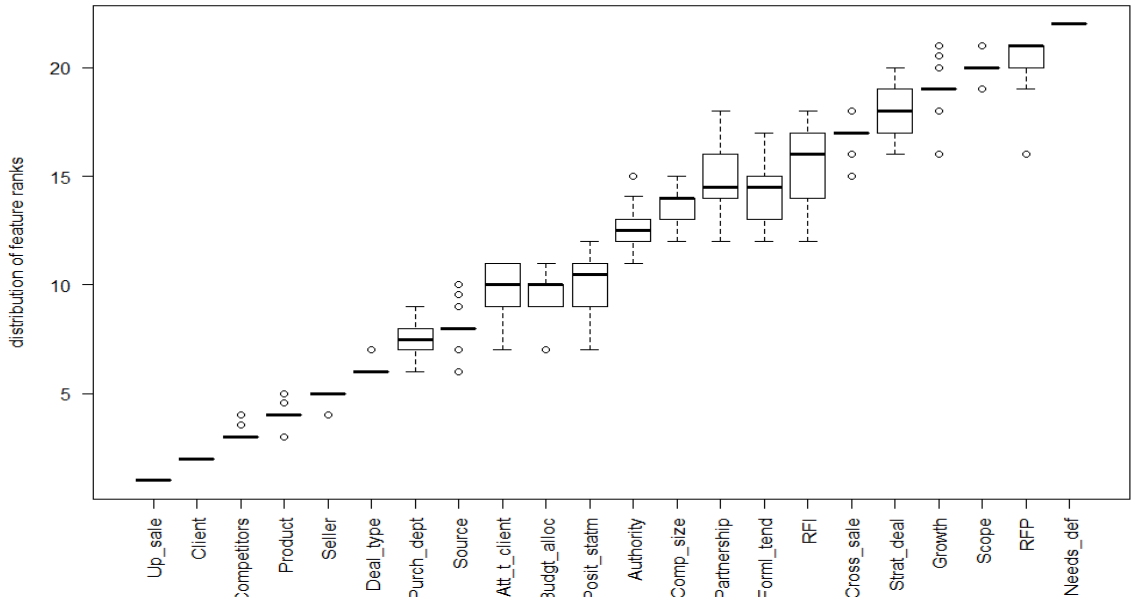
Data set and ground truth

In this section, we introduce the data set and ground truth for feature ranks. We try to identify the median rank of a particular feature obtained from the random samples of size $|V|$. We use median instead of mean to obtain robust results.

As a use case we use a real world B2B sales data set (Bohanec, 2017) with 448 instances, 22 features and a class variable with two values. To form an optimization problem we need ground truth ranks of features which, for practical problems, are unavailable. We estimate the ground truth ranks of features (a_1,...,a_t), t being the number of features, we rank the features with a selected feature ranking algorithm on the complete data set using 10-fold cross-validation. In this paper, we use ReliefF feature evaluation (Robnik-Šikonja et al., 2003), known for its robustness and ability to detect strongly dependent features. Figure 1 shows box-and-whiskers plots for all 22 attributes. The ranks of the most important features are stable, as indicated by low

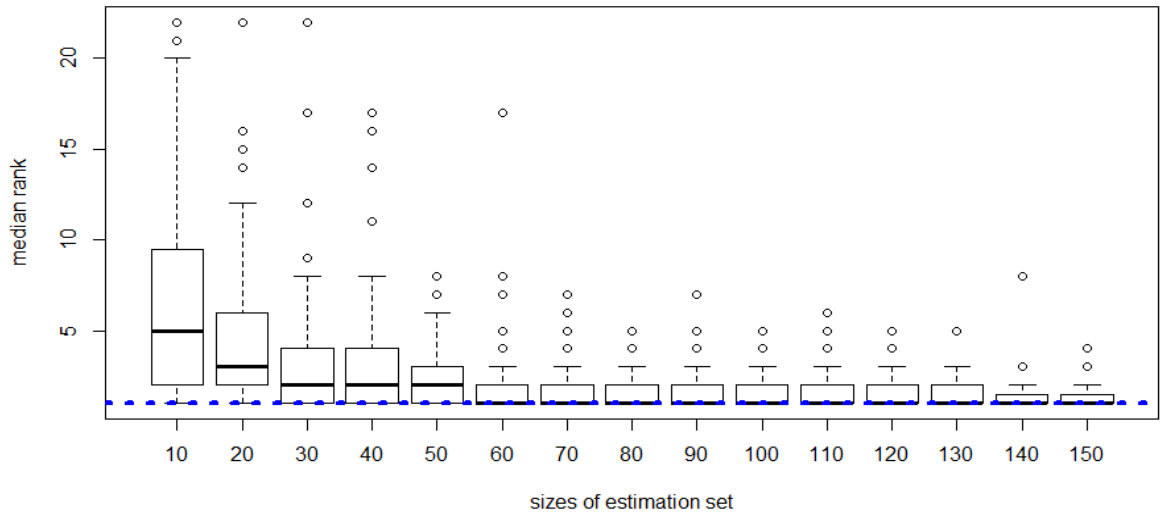
variance around median in box-and-whiskers plots. Similarly, the least performing features are consistently the last.

Figure 1
Feature ranks on the complete data set (ground truth), estimated with ReliefF using 10-fold cross-validation. Horizontal axis shows features and vertical axis shows distribution of their ReliefF ranks.



Source: Authors' work

Figure 2
Distribution of ReliefF ranks for feature "Up_sale" ranked 1st for different sizes of estimation set (sampled directly from the full data set). Dotted blue line indicates true rank.

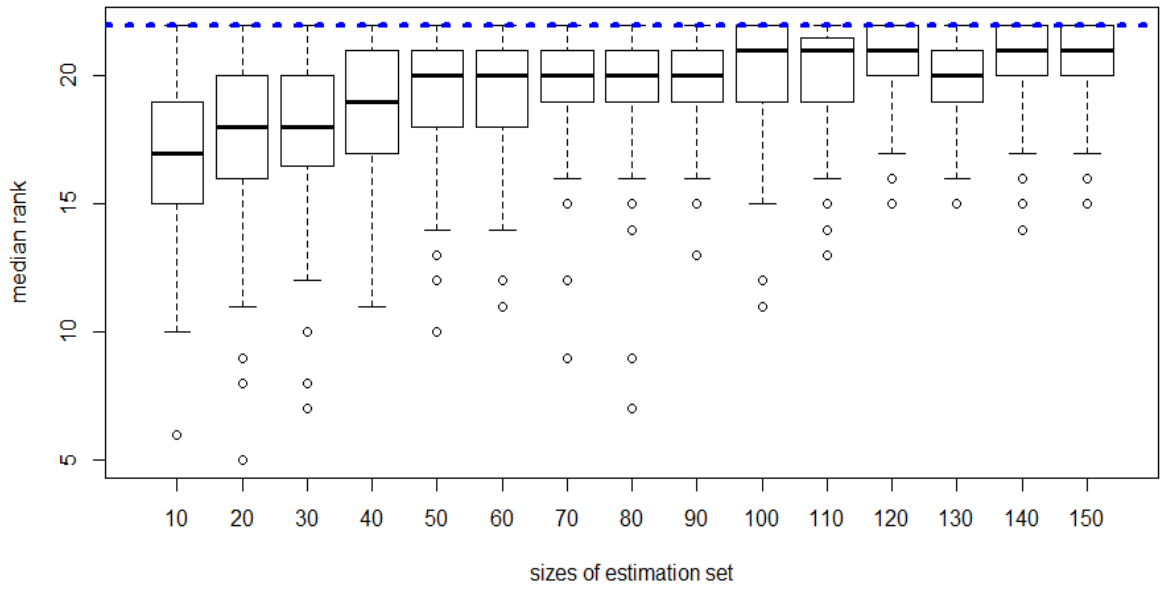


Source: Authors' work

Next, we observe feature ranks as the number of instances increases. We start with a random sample of size 10 and increase the sample size to 150 in increments of 10. Each sample size is resampled 100-times from a complete data set. The distributions of obtained ranks for the feature with the strongest impact from Figure 1 (i.e. "Up_sale") for different sample sizes are reported in Figure 2. We see that this feature is consistently ranked among the best features even with very low number of instances. From sample size 30 this feature is ranked among top 5 features with high probability.

The rank distributions of the least performing feature "Needs_def" (ranked 22nd in Figure 1) are presented in Figure 3. The results show that this feature indicates a clear tendency to bottom ranks from the smallest subset size on. From sample size 10, the vast majority of obtained ranks are larger than 10, as indicated by the notch of the box-and-whiskers plot.

Figure 3
Distribution of ReliefF ranks for feature "Needs_def" ranked 22nd (last) for different sizes of estimation set (sampled directly from the full data set). Dotted blue line indicates true rank.



Source: Authors' work

Formalization of the problem

We assume that we estimate features' impact within an existing data set. We evaluate the number of instances needed for a feature to reliably show its impact given that the ground truth is known.

Therefore, our goal is to find the smallest size of a random subset of instances $|V|$, which assures that for a given feature a_i, ranked by function R, the rank of the feature computed on V is close to the the rank obtained on the complete data set:

$$|R(a^{V}_{i}) - R(a_i)| \leq \varepsilon \quad (1)$$

We use the following notation: R is a ranking function, a_i represents feature i, $R(a_i)$ is the rank of feature a_i on the complete data set of size n, $R(a_i^V)$ is rank of feature a_i on the subset V of the data set, and $\varepsilon \geq 0$ is the tolerance of ranking error.

Eq. (1) determines the minimal size of a data set that assures with high probability (at least τ) that the rank of a given feature a_i is close to its true rank. We approximate the true ranks by ranking features on the complete data set.

$$|V|_i^{min} = \arg \min_{|V| \in [1,n]} [P(|R(a_i) - R(a_i^V)| \leq \varepsilon) \geq \tau] \quad (2)$$

For example, if we set $\tau = 0.95$, we expect that with 95% probability we will not make error larger than ε when estimating feature a_i from sample of size $|V|_i^{min}$ instead of from the complete data set. We expect to find sample sizes $|V|_i^{min}$ which will be robust to the variations in the randomly sampled training data for the given feature and the selected ranking function R. Discussion on stability of feature evaluation can be found in (Kalousis et al., 2007). We use bootstrap sampling (Kohavi, 1995; Davison et al., 1997) to obtain confidence intervals (CI) for determination of $|V|_i^{min}$.

For practical use we propose two variants of Eq. (1). We are interested if a given feature might be useful in a predictive model, in this case its rank has to be lower than a prespecified rank threshold L. On the other hand, we are also interested if a given feature can be safely discarded from further consideration. In this case its rank has to be higher than a threshold H. Both cases are formalized below in Eqs. (2) and (3) and can be estimated with ranking function R applied to bootstrap samples.

$$|V|_i^L = \arg \min_{|V| \in [1,n]} [P(|R(a_i^V) \leq L) \geq \tau] \quad (3)$$

$$|V|_i^H = \arg \min_{|V| \in [1,n]} [P(|R(a_i^H) \leq H) \geq \tau] \quad (4)$$

Experiments

The aim of our study is to show a practical method how to estimate the number of instances needed for a new feature to reliably estimate its rank within an existing set of features and existing data set. Our procedure is as follows. For each feature we gradually increase the sample size $|V|$, randomly select a sample of this size from the full data set 30-times, and bootstrap each sample 500 times. The bootstraped samples are used with ranking function ReliefF and form a basis to calculate the median and confidence interval (CI) for each size. The collection of these estimates is illustrated with pseudo code in Algorithm 1. Actual experiments are run within R environment using libraries *caret* (Kuhn, 2017), *CORElearn* (Robnik-Šikonja et al., 2017) and *ggplot2* (Wickham, 2009).

Results on a sales forecasting problem

In practice, users are providing instances of data in small chunks. In order to estimate feature impact we can use only these instances. To account for variance in the obtained sample provided by users we use bootstrap confidence interval estimation that uses sampling with replacement.

First, we analyze features in the existing data set to see what we can expect for new features. We are particularly interested in top performing features (which we

want to retain) and least performing features (which we can safely discard). Our testing data set contains 22 features. Based on previous research (Bohanec et al., 2015a, Figure 2), we know that 8 features can be sufficient for random forest classifier to reach satisfactory performance, therefore we set the threshold L to 11 (incorporating a safety band of 3 (this would correspond to $\varepsilon = 3$ in Eq. 1)). We set the threshold for discarding the highest ranking features to 15. Results of experiments produce figures similar to Figures 2 and 3. From the distributions depicted with box-and-whiskers plots we can even visually determine how many instances are required to reliably recognize top ranked feature's and how many to reliably discard the features with high ranks.

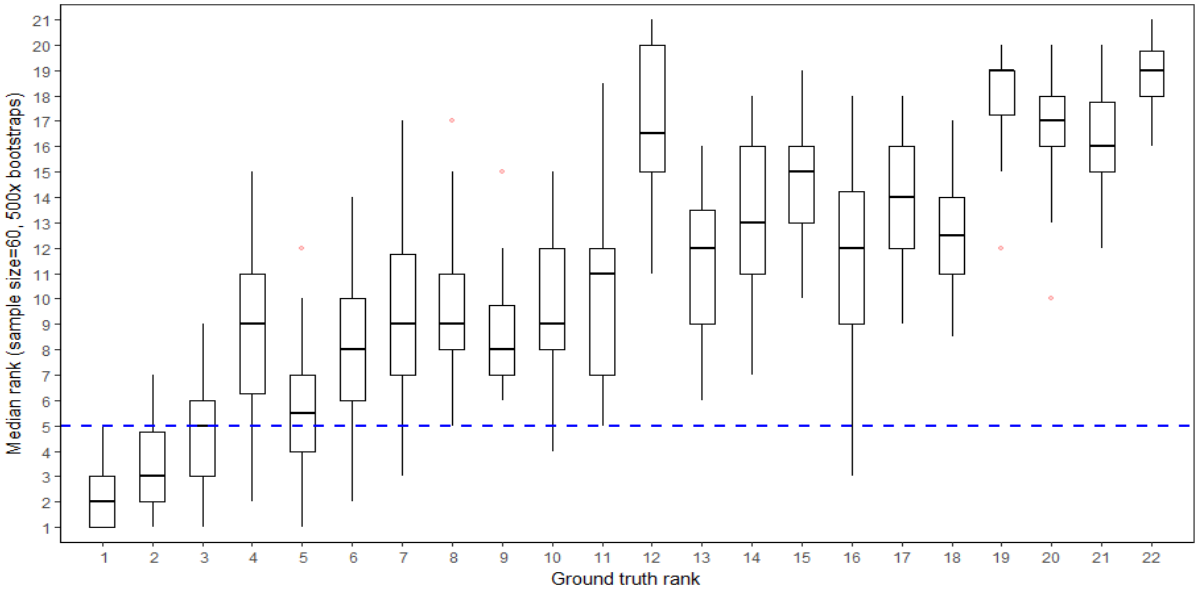
Algorithm 1
Distribution of feature ranks for different number of instances

```
1: procedure SubsetSizes(parameters: data, numExperiments, initialSize, step )
2:     subsetSize = initialSize
3:     while subsetSize ≤ size(data) do
4:         for q in 1:30 do
5:             sampleData = Sample(data, subsetSize, replace = FALSE)
6:             for k in 1: 500 do
7:                 trialData = BootstrapData(sampleData, subsetSize, replace = TRUE)
8:                 trialRanks[k] = ReliefF(trialData)          ◦ get ranks for all features
9:             end for
10:          medianRanks[q] = median(trialRanks)          ◦ compute median ranks for all features
11:        end for
12:        Store medianRanks for current subsetSize
13:        subsetSize = subsetSize + step
14:    end while
15:   Return stored rank distributions for all sample sizes
16: end procedure
```

Source: Authors' work

Figure 4 shows results we obtained for all existing features. We simulated a scenario where a new feature described with 60 instances is provided. Based on that we can provide the following guidelines for a user with a given number of available instances describing a new feature. To estimate feature's rank, perform 500 repetitions of bootstrap sampling and feature evaluation with ReliefF. The median rank from bootstrap repetitions shall be recorded and compared with rank distributions of existing features using the same number of instances. Figure 4 gives an example of rank distributions for 60 instances. E.g., if the median rank of a new feature would be 5, the horizontal line passing the rank 5 reveals which features exhibited similar behavior with this number of instances. Based on that one can take one of the three decisions: a) if the obtained rank line crosses distributions of mostly top ranked features, retain the feature and use it in the model from that time onwards, b) if the rank's line crosses mostly distributions of least ranked features, discard the feature, or c) if neither a) or b) is true, postpone the decision and try to collect more data (depending on the effort, cost of data collection, etc.).

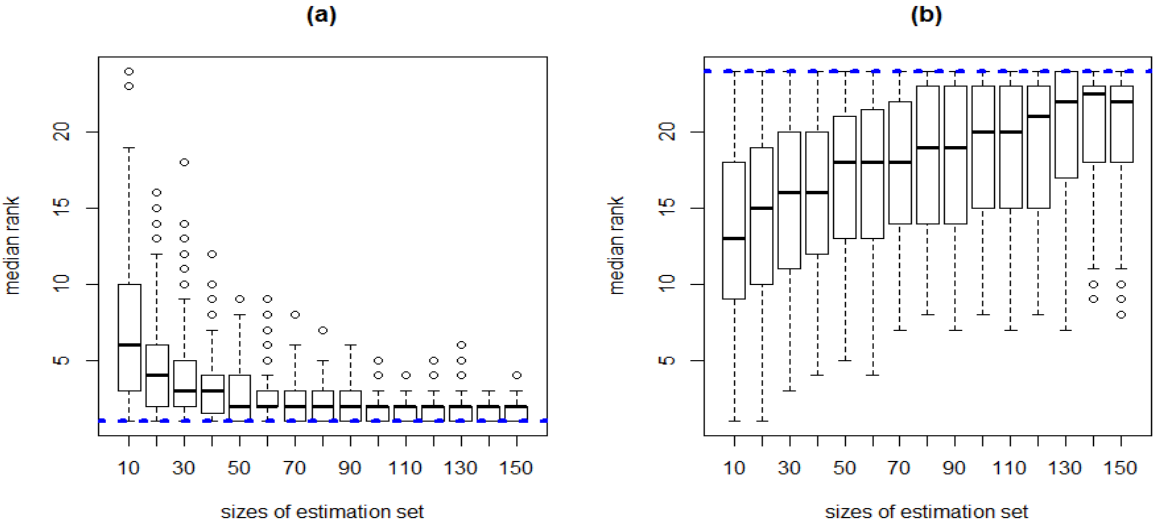
Figure 4
Rank of all features, based on sample size 60 with 500 bootstraps, repeated 30-times



Source: Authors' work

Two UCI data sets

Figure 5
On CKD data set with missing values true ranks are not reached, neither for (a) top ranked feature nor for (b) the bottom ranked feature.

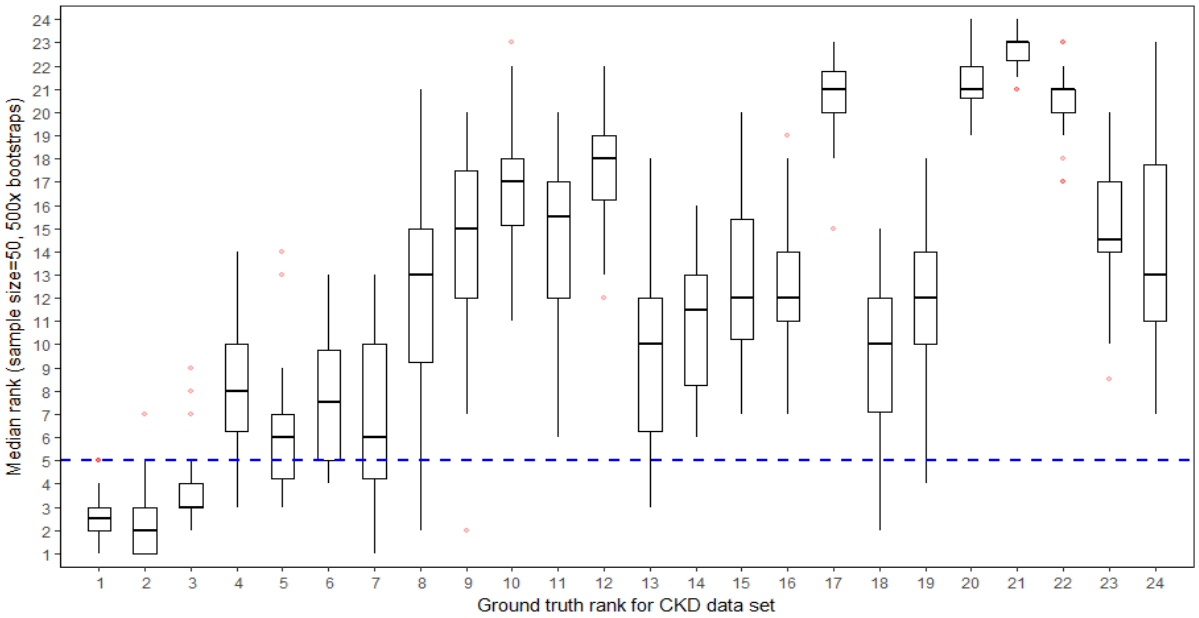


Source: Authors' work

To strengthen our analysis and generalize the conclusions, we applied presented approach to two publicly available data sets from UCI Machine Learning Repository (Lichman, 2013). The Wine data set (Forina et al., 1991) has 12 features + class variable and 178 samples without missing values. The results are similar to the results on B2B dataset, therefore we omit further discussion. The Chronic Kidney Disease data set (Soundarapandian, 2015) has 24 features + class variable and 400 samples,

with total 1012 missing values, 10.5% of all values. For this case, Figure 5 shows that both the top (a) and the bottom (b) ranked feature converges toward a ground truth rank (dotted blue line) but don't reach it even with a sample size of 150 samples. Figure 6 shows that only the top three ranked features reveal their impact with small sample size (50 in this case), the rest of the features display high volatility. We assume that this might be caused by missing values; however, this needs to be further researched.

Figure 6
Ranks of all CKD features, based on sample size 50 with 500 bootstraps.



Source: Authors' work

Conclusions

We address the problem of updates to the existing classification models as a result of changed problem understanding of domain experts. Experts consider adding various new features to the classification model and are interested to assess their potential impact with minimal data collection effort. For this purpose, we formalize a problem of minimal number of instances needed to reliably estimate the impact of new features added to the existing data set. We use the existing data set as a proxy for ground truth ranks.

The results on the analyzed B2B data set show that relatively low number of instances is required to determine impacts of top performing features and least performing features. Such results are promising for practical use and indicate that a reasonably low effort of B2B practioners is required to assess the impact of useful new features. The results on the additional Wine data set show similar trends as the B2B data set. The results on CKD data set show similar trends, but exhibit higher volatility that may be result of many missing values present in this data set.

In the future, our approach shall be tested in several domains of various character to draw more general conclusions about the minimum number of required instances. In addition, the impact of missing values on the stability of feature ranks requires further research. A possible direction would use synthetic data sets with known characteristics to better control the information content and volatility of feature ranks due to missing values.

References

1. Beleites, C., Neugebauer U., Bocklitz T., Krafft, C., Popp, J. (2013), "Sample size planning for classification models", Analytica Chimica Acta, Vol. 760, pp. 25-33.
2. Bohanec, M. (2017), "A public B2B data set used for qualitative sales forecasting research", available at: http://www.salvirt.com/research/B2Bdataset/ (01 August 2017).
3. Bohanec, M., Kljajić Borštnar, M., Robnik-Šikonja, M. (2015a), "Feature subset selection for B2B sales forecasting", in Zadnik Stirn L., Žerovnik J., Kljajić Borštnar M., Drobne S. (Eds.), 13th International Symposium on Operational Research, SDI-SOR, Bled, Slovenia, pp. 285-290.
4. Bohanec, M., Kljajić Borštnar, M., Robnik-Šikonja, M. (2015b), "Machine learning data set analysis with visual simulation", in Kljajić L., Lasker G. E. (Eds.), Advances in simulation-based decision support & business intelligence, Vol. 5, Tecumseh: International Institute for Advanced Studies in Systems Research and Cybernetics, Baden-Baden, Germany, pp. 16-20.
5. Bohanec, M., Kljajić Borštnar, M., Robnik-Šikonja, M. (2016), "Sample size for identification of important attributes in B2B sales", in Scitovski R., Zekić-Sušac M. (Eds.), 16th International Conference on Operational Research, CRORS, Osijek, Croatia, p. 133.
6. Davison, A. C., Hinkley, D. V. (1997), Bootstrap methods and their application, Vol. 1, Cambridge University Press.
7. Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., Ngo, L. H. (2012), "Predicting sample size required for classification performance", BMC medical informatics and decision making, Vol. 12, No. 1, pp. 1-8.
8. Forina, M. et al. (1991), "UCI machine learning repository - using chemical analysis determine the origin of wines", available at: https://archive.ics.uci.edu/ml/datasets/Wine (01 January 2018).
9. Guyon, I., Elisseeff, A. (2003), "An introduction to variable and feature selection", Journal of machine learning research, Vol 3, No. 1, pp. 1157-1182.
10. Kalousis, A., Prados, J., Hilario, M. (2007), "Stability of feature selection algorithms: a study on high-dimensional spaces", Knowledge and information systems, Vol. 12, No. 1, pp. 95-116.
11. Kohavi R. (1995), "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", in Mellish, C. S. (Ed.), Artificial Intelligence Proceedings 14th International Joint Conference, Morgan Kaufmann, USA, pp. 1137-1145.
12. Kuhn, M. (2017), "A short introduction to the caret package", available at: https://cran.rproject.org/web/packages/caret/vignettes/caret.pdf (01 August 2017).
13. Lichman, M. (2013), "UCI Machine Learning Repository", available at: http://archive.ics.uci.edu/ml (01 February 2018).
14. Robnik-Šikonja, M., Kononenko, I. (2003), "Theoretical and empirical analysis of ReliefF and RReliefF", Machine learning, Vol. 53, No.1-2, pp. 23-69.
15. Robnik-Šikonja, M., Savicky, P. (2017), "CORElearn - classification, regression, feature evaluation and ordinal evaluation", R package version 1.51.2.
16. Soundarapandian, P. (2015), "UCI machine learning repository - the chronic kidney disease prediction data set", available at: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease (01 January 2018).
17. Wickham, H. (2009), ggplot2: Elegant Graphics for Data Analysis, Springer, New York.

About the authors

Marko Bohanec received his Ph.D. in Management Information Systems from the University of Maribor. He received his MSc from Faculty of Economy and BSc from Faculty of Computer and Information Science, both University of Ljubljana, Slovenia. Professionally, he consults companies how to minimize risks in their sales performance and staff development. His research interests relate to improvements in business management by utilizing advancements in the field of machine learning. He is author of several scientific articles published in conferences and international journals including Expert Systems with Applications and Industrial Management & Data Systems. The author can be contacted at **marko.bohanec@salvirt.com.**

Mirjana Kljajić Borštnar received her Ph.D. in Management Information Systems from the University of Maribor. She is an Associate Professor at the Faculty of Organizational Sciences, University of Maribor and a member of Laboratory for Decision Processes and Knowledge-Based Systems. Her research work covers decision support systems, multi-criteria decision-making, system dynamics, data mining, and organizational learning. She is the author and co-author of several scientific articles published in recognized international journals and conferences, including Group Decision and Negotiation and System Dynamics Review. The author can be contacted at **mirjana.kljajic@fov.uni-mb.si**.

Marko Robnik-Šikonja received his Ph.D. in computer science and informatics in 2001 from the University of Ljubljana. He is Professor of Computer Science and Informatics and Head of Artificial Intelligence Chair at the University of Ljubljana, Faculty of Computer and Information Science. His research interests include machine learning, data and text mining, knowledge discovery, network mining, and their practical applications. He is a (co)author of more than 100 publications in scientific journals and international conferences and maintains three open-source analytic tools. The author can be contacted at **marko.robniksikonja@fri.uni-lj.si**.