View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Crossref

Comparison of Cross-Validation and Test Sets Approaches to Evaluation of Classifiers in Authorship Attribution Domain

Grzegorz Baron(✉)

Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
grzegorz.baron@polsl.pl

Abstract. The presented paper addresses problem of evaluation of decision systems in authorship attribution domain. Two typical approaches are cross-validation and evaluation based on specially created test datasets. Sometimes preparation of test sets can be troublesome. Another problem appears when discretization of input sets is taken into account. It is not obvious how to discretize test datasets. Therefore model evaluation method not requiring test sets would be useful. Cross-validation is the well-known and broadly accepted method, so the question arose if it can deliver reliable information about quality of prepared decision system. The set of classifiers was selected and different discretization algorithms were applied to obtain method invariant outcomes. The comparative results of experiments performed using cross-validation and test sets approaches to system evaluation, and conclusions are presented.

1 Introduction

Evaluation of classifier or classifiers applied in a decision system is the important step during a model building process. Two approaches are typical: cross-validation and using of test datasets. Both have some advantages and disadvantages. Cross-validation is easy to apply and in different application domains is accepted as good tool for measuring of classifiers performance. Evaluation based on test datasets requires at the beginning preparation of special sets containing data disjunctive of training one used during the creation process of a decision system. Sometimes it can be difficult to satisfy such condition.

Another issue, which arose during the author's former research, was utilization of test sets in conjunction with discretization of input data [3]. There are fundamental questions, how discretize test datasets in relation to learning sets to keep both sets coherent. Some approaches were analyzed, but they did not deliver unequivocal results. Therefore another idea came out - use of cross-validation instead of test data to validate the decision system. Such approach required deeper investigation and comparison with the first method of model validation. The paper presents experimental results, discussion and conclusions about that issue.

Authorship attribution is a part of stylometry which deals with recognition of texts' authors. Subject of analysis ranges from short Twitter messages to huge

© The Author(s) 2016
T. Czachórski et al. (Eds.): ISCIS 2016, CCIS 659, pp. 81–89, 2016.
DOI: 10.1007/978-3-319-47217-1_9

works of classical writers. Machine learning techniques and statistic-oriented methods are mainly involved in that domain. Different authorship attribution tasks have been categorized in [12], and three kinds of problems were formulated: profiling – there is no candidate proposed as an author; the needle-in-a-haystack – author of analyzed text should be selected from thousands of candidates; verification – there is an candidate to be verified as author of text.

The first important issue is to select characteristic features (attributes) to obtain author invariant input data which ensure good quality and performance of decision system [16]. Linguistic or statistical methods can be applied for that purpose. The analysis of syntactic, orthographic, vocabulary, structure, and layout text properties can be performed in that process [9].

The next step during building a decision system for authorship attribution task is selecting and applying the classifier or classifiers. Between different methods some unsupervised ones like cluster analysis, multidimensional scaling and principal component analysis can be mentioned. Supervised algorithms are represented by neural networks, decision trees, bayesian methods, linear discriminant analysis, support vector machines, etc. [9,17]

As aforementioned the aim of presented research was to compare two general approaches to evaluation of decision system: cross-validation [10] and test datasets utilization. To obtain representative results, a set of classifiers was chosen, applied and tested for stylometric data performing authorship attribution tasks. The idea was to select classifiers characterized by different ways of data processing. Finally the following suite of classifiers was applied: Naive Bayes, decision tree C4.5, k-Nearest Neighbors k-NN, neural networks – multilayer perceptron and Radial Basis Function network – RBF, PART, Random Forest. Test were performed for non-discretized and discretized data applying different approaches to test datasets discretization [3].

The paper is organized as follows. Section 2 presents the theoretical background and methods employed in the research. Section 3 introduces the experimental setup, datasets used and techniques employed. The test results and their discussion are given in Sect. 4, whereas Sect. 5 contains conclusions.

2 Theoretical Background

The main aims of presented research were analysis and comparison of cross-validation and test dataset approaches to evaluation of classifier or classifiers used in decision system especially in authorship attribution domain. Therefore a suite of classifiers has been set. The main idea was to select classifiers which behave differently because of performed algorithm and way of data processing. The final list of used classifiers contains: decision trees – PART [6] and C4.5 [14], Random Forest [4], k-Nearest Neighbors [1], Multilayer Perceptron, Radial Basis Function network, Naive Bayes [8].

Discretization is a process which allows to change the nature of data – it converts continuous values into nominal (discrete) ones. Two main circumstances can be mentioned, where discretization may or even must be applied. The first

situation is when there are some suspicions about possible improvement of a decision system quality when discretized data is applied [2]. The second one is when method or algorithm employed in decision system can operate only on nominal, discrete data.

Because discretization reduces amount of data to be processed in a subsequent modules of decision system, sometimes it allows to filter information noise or allow to represent data in more consistent way. But on the other hand improper discretization application can lead to significant loss of information, and to degradation of overall performance of decision system.

Discretization algorithms can be divided basing on the different criterions. There are global methods which operate on whole attribute domain or local ones which process only part of input data. There are supervised algorithms which utilize class information in order to select bin ranges more accurately or unsupervised ones which perform only basic splitting of data into desired number of intervals [13]. Unsupervised methods are easier in implementation but supervised ones are considered to be better and more accurate.

In the presented research four discretization methods were used: equal width binning, equal frequency binning, as representatives of unsupervised algorithms, and supervised Fayyad & Irani's MDL [5] and Kononenko MDL [11].

The equal width algorithm divides the continuous range of a given attribute values into required number of discrete intervals and assigns to each value a descriptor of appropriate bin. The equal frequency algorithm splits the range of data into a required number of intervals so that every interval contains the same number of values.

During the developing of decision system, where input data is discretized and classifier is evaluated using test datasets, another question arises, namely how to discretize test datasets in relation to training data. Depending on the discretization methods different problems can appear such as uneven number of bins in training and test data, or cut-points which define boundaries of bins can be different in both datasets. That can lead to some inaccuracy during the evaluation of decision system. In [3] three approaches to discretization of test datasets were proposed:

- "independent" (Id) – training and test datasets are discretized separately,
- "glued" (Gd) – training and test datasets are concatenated, the obtained set is discretized, and finally resulting dataset is split back into learning and test sets,
- "test on learn" (TLd) – firstly training dataset is discretized, and then test set is processed using cut-points calculated for training data.

3 Experimental Setup

The following steps were performed during the execution of experiments:

1. training and test data preparation,
2. discretization of input data applying selected algorithms using various approaches to test data processing,

3. training of selected classifiers,
4. system evaluation using cross-validation and test data approaches.

Input datasets were built basing on the several works of two male and two female authors. To obtain input data containing characteristic features satisfying author invariant requirement the following procedure was employed. Some linguistic descriptors from lexical and syntactic groups were chosen [15]. The works of each author were divided into parts. Then for each part frequencies of usages of selected attributes were calculated. Finally separate training and test sets were prepared with two classes (corresponding to two authors) in each. Attention was given during data preparation in order to obtain well-balanced training sets.

All experiments were performed using WEKA workbench, especially discretization methods and classifiers come from that software suite. It was necessary to make some modifications and develop additional methods to implement discretization algorithms allowing to discretize test data in "test on learn" and "glued" manner. Unsupervised discretization such as equal width and equal frequency were performed for required number of bins parameter ranged from 2 to 10. Base on the author's former experiences that was the range, where results are worth of notice.

According to the main aim of the presented research classifiers were evaluated using cross-validation and test datasets. Cross-validation was performed typically in 10-folds version. As a measure of classifier quality the number of correctly classified instances was taken.

4 Results and Discussion

The experiments were performed separately for male and female authors but final results were averaged for analysis and presentation purposes. For both neural network classifiers the best results obtained during experiments performed using multistart strategy are presented. Abbreviations used for classifiers naming in Figs. 1–3 are as follows: NB – Naive Bayes, C4.5 – decision tree C4.5, Knn – k-Nearest Neighbors, PART – decision tree PART, RF – Random Forest, RBF – Radial Basis Function network, MLP – Multilayer Perceptron. Additionally in Fig. 3 postfix "_T" denotes results obtained for evaluation using test data whereas postfix "_CV" is used for cross-validation results.

Results of the preliminary experiments performed for non-discretized data are presented in Fig. 1. It is easy to notice that classifiers performance measured using cross-validation are about 10 % better than results obtained for evaluation performed using test datasets. Only k-Nearest Neighbor classifier behave slightly better for evaluation using test data.

Figure 2 shows comparative results obtained for both analyzed evaluation approaches for data discretized using Kononenko MDL and Fayyad & Irani MDL respectively. Because test datasets were discretized using "Test on Learn", "Glued", and "Independent" approaches, the X axis is parted into three sections

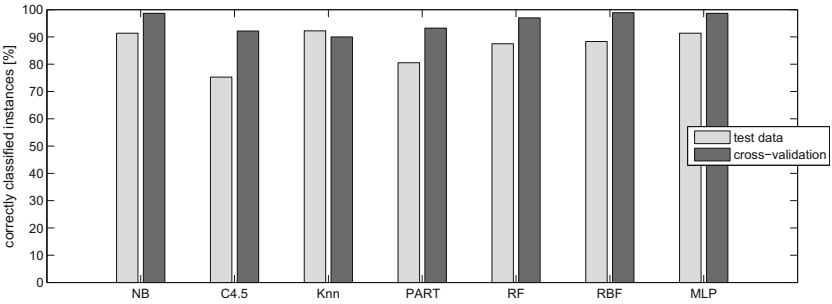

Fig. 1. Performance of classifiers for non-discretized data for evaluation performed using cross-validation and test datasets

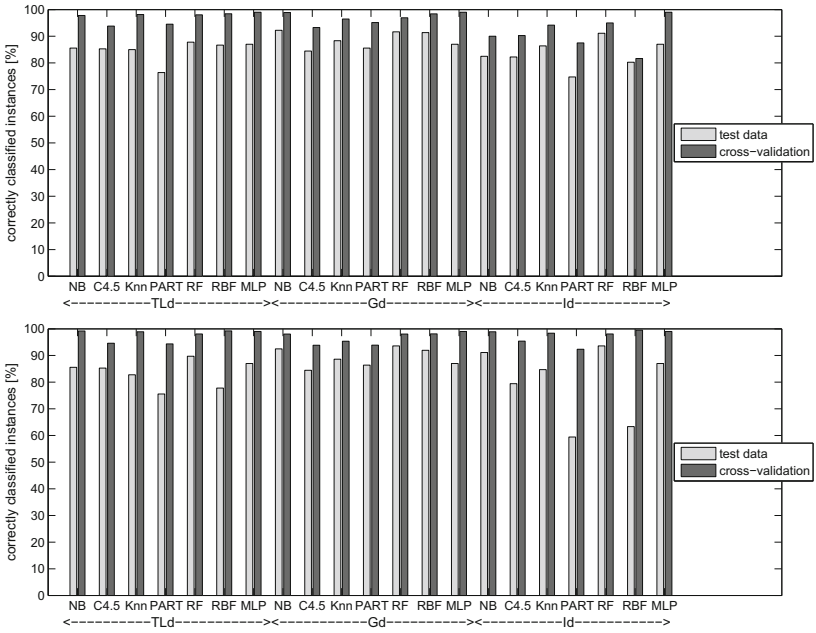

Fig. 2. Performance of classifiers for data discretized using supervised Kononenko MDL (above) and Fayyad & Irani MDL (below) for evaluation performed using cross-validation and test datasets. Three sections of the X axis present evaluation results obtained for test datasets discretized using "Test on Learn" – TLd, "Glued" – Gd, and "Independent" – Id approaches

which present results for mentioned ways of discretization. The huge domination of outcomes obtained for cross-validation evaluation is visible. Especially for "Independent" discretization of test datasets differences are big for PART and RBF classifiers.

Results obtained for unsupervised equal width and equal frequency discretization are shown in Fig. 3. Because experiments were parametrized using required number of bins ranged from 2 to 10, the boxplot diagrams were used to clearly visualize averaged results and relations between cross-validation and test set approaches to classifiers evaluation. The general observations are similar to the previous ones. For all classifiers, for all ways of discretization of test sets, and for both equal width and equal frequency discretization methods number

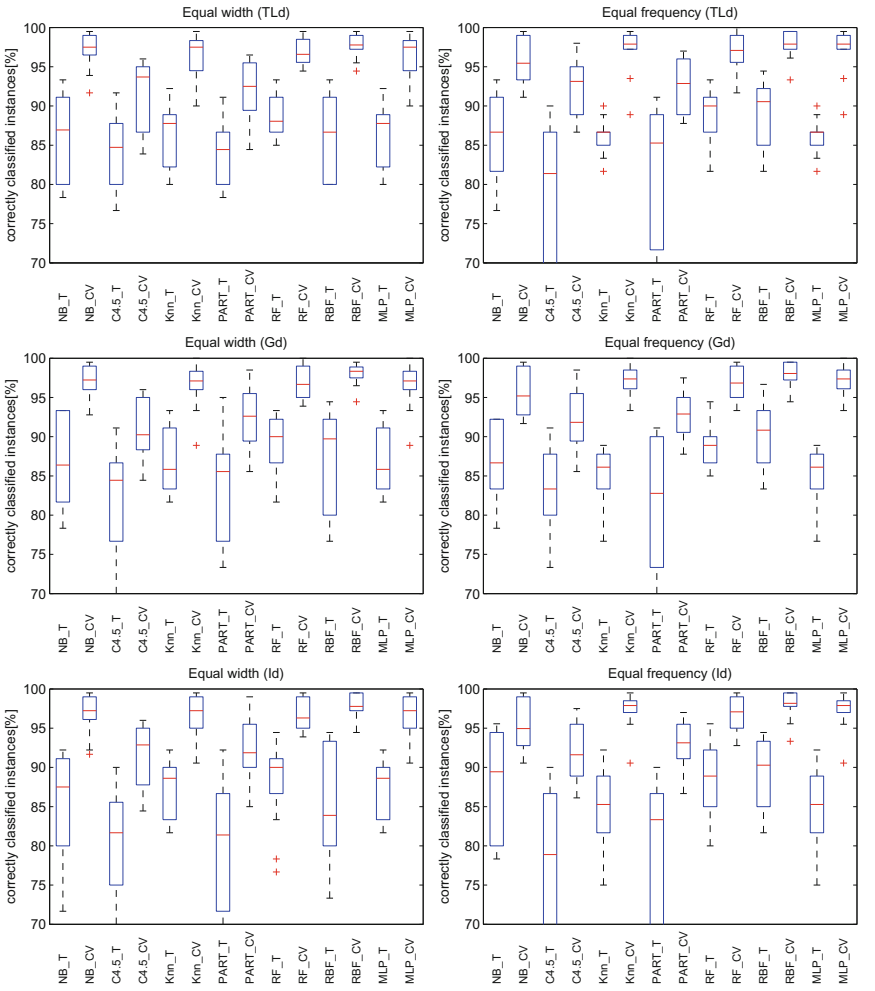

Fig. 3. Performance of classifiers for data discretized using unsupervised equal width (left column) and equal frequency (right column) discretization performed using the following approaches: "Test on Learn" – TLd (top row), "Glued" – Gd (middle row), and "Independent" – Id (bottom row), for evaluation performed using cross-validation ("_CV") and test datasets ("_T")

of correctly classified instances reported for cross-validation evaluation is bigger than for test dataset approach. The average difference is about 10 % (taking the medians of boxplots as reference points).

Summarizing the presented observations it can be stated that for almost all experiments (only one exception was observed) evaluation performed using cross-validation delivered quality measurements about 10 % greater comparing to the evaluation based on test datasets. In some cases that results reached 100 %. This is a problem because can lead to false conclusions about real quality of created decision system. Practically it is impossible to develop a system working with so high efficiency. Evaluation based on test datasets proved this opinion. Test sets were prepared basing on the texts other than that used for training of classifiers. So that evaluation results can be considered as more reliable. Depending on the classifier and discretization method they are smaller up to 30 %.

The general conclusion is that cross-validation which is acceptable and broadly used in different application domains is rather not useful for evaluating of decision systems in authorship attribution tasks performed in conditions and for data similar to that presented in the paper. If one decides to apply this method, must take into account that real performance of the system is much worse than reported using cross-validation evaluation.

5 Conclusions

The paper presents research on evaluation of decision systems in authorship attribution domain. Two typical approaches, namely cross-validation and evaluation based on specially created test datasets are considered. The research was the attempt to answer the question if evaluation using test datasets can be replaced by cross-validation to obtain reliable information about overall decision system quality. The set of different classifiers was selected and different discretization algorithms were applied to obtain method invariant outcomes. The comparative results of experiments performed using cross-validation and test sets approach to system evaluation are shown.

For almost all experiments (there were only one exception) evaluation performed using cross-validation delivered quality measurements (percent of correctly classified instances) about 10 % greater comparing to the evaluation based on test datasets. There were outliers where difference up to 30 % could be observed. On the other hand in some cases number od correctly classified instances for cross-validation was equal to 100 % what is not probable in real live tasks.

Concluding the research, it must be stated that cross-validation is rather not useful method for evaluating of decision systems in authorship attribution domain. It can be conditionally applied but strong tendency to overrating the quality of examined decision system must be taken into consideration.

Acknowledgments. The research described was performed at the Silesian University of Technology, Gliwice, Poland, in the framework of the project BK/RAu2/2016. All experiments were performed using WEKA workbench [7] basing on texts downloaded from http://www.gutenberg.org/.

Open Access. This chapter is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, duplication, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, a link is provided to the Creative Commons license and any changes made are indicated.

The images or other third party material in this chapter are included in the work's Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work's Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt or reproduce the material.

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. In: Machine Learning, pp. 37–66 (1991)
2. Baron, G.: Influence of data discretization on efficiency of Bayesian Classifier for authorship attribution. Procedia Comput. Sci. **35**, 1112–1121 (2014)
3. Baron, G., Harezlak, K.: On Approaches to discretization of datasets used for evaluation of decision systems. In: Czarnowski, I., Caballero, A.M., Howlett, R.J., Jain, L.C. (eds.) Intelligent Decision Technologies 2016, vol. 57, pp. 149–159. Springer, Cham (2016)
4. Breiman, L., Schapire, E.: Random forests. In: Machine Learning, pp. 5–32 (2001)
5. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuousvalued attributes for classification learning. In: 13th International Joint Conference on Articial Intelligence, vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)
6. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization, pp. 144–151. Morgan Kaufmann (1998)
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
8. John, G., Langley, P.: Estimating continuous distributions in bayesian classifiers. In. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345. Morgan Kaufmann (1995)
9. Juola, P.: Authorship attribution. Found. Trends Inf. Retrieval **1**(3), 233–334 (2008)
10. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence, pp. 1137–1143 (1995)
11. Kononenko, I.: On biases in estimating multi-valued attributes. In: 14th International Joint Conference on Articial Intelligence, pp. 1034–1040 (1995)
12. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inform. Sci. Technol. **60**(1), 9–26 (2009)
13. Kotsiantis, S., Kanellopoulos, D.: Discretization techniques: a recent survey. Int. Trans. Comput. Sci. Eng. **1**(32), 47–58 (2006)

14. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
15. Stańczyk, U.: Ranking of characteristic features in combined wrapper approaches to selection. Neural Comput. Appl. **26**(2), 329–344 (2015)
16. Stańczyk, U.: Establishing relevance of characteristic features for authorship attribution with ANN. In: Decker, H., Lhotská, L., Link, S., Basl, J., Tjoa, A.M. (eds.) DEXA 2013, Part II. LNCS, vol. 8056, pp. 1–8. Springer, Heidelberg (2013)
17. Stańczyk, U.: Rough set and artificial neural network approach to computational stylistics. In: Ramanna, S., Howlett, R.J. (eds.) Emerging Paradigms in ML and Applications. SIST, vol. 13, pp. 441–470. Springer, Heidelberg (2013)