

Random Forest Feature Selection for Data Coming from Evaluation Sheets of Subjects with ASDs

Krzysztof Pancierz, Wiesław Paja
University of Rzeszów, Poland
Email: {kpancerz,wpaja}@ur.edu.pl

Jerzy Gomuła
Cardinal Stefan Wyszyński University
Warsaw, Poland
Email: jerzy.gomula@wp.pl

Abstract—We deal with the problem of initial analysis of data coming from evaluation sheets of subjects with Autism Spectrum Disorders (ASDs). In our research, we use an original evaluation sheet including questions about competencies grouped into 17 spheres. In the paper, we are focused on a feature selection problem. The main goal is to use appropriate data to build simpler and more accurate classifiers. The feature selection method based on random forest is used.

I. INTRODUCTION

AUTISM is a brain development disorder that impairs social interaction and communication, and causes restricted and repetitive behaviors. Autism spectrum disorders can dramatically affect a child's life, as well as that of their families, schools, friends and the wider community.

The main aim of our research is to adapt computational intelligence methods for computer-aided decision support in diagnosis and therapy of persons with ASDs. In the first step of our research, we are interested in initial analysis of data coming from evaluation sheets of subjects with ASDs. The evaluation sheet, we use in the research, is an original sheet including questions (more than 300) about competencies of the subjects grouped into 17 spheres, among others, self-service, communication, cognitive, physical, as well as the sphere responsible for functioning in the social and family environment.

An initial analysis is focused on the data preprocessing step. The preprocessed data can be used to build simpler and more accurate classifiers. It is obvious, that an increasing number as well as complexity of classification rules make it difficult to be validated by domain experts. Experiments showed that in case of our evaluation sheet, over 300 features corresponding to questions (even divided into spheres) lead to less accurate classifiers with complex classification rules. Therefore, there is an important problem to select appropriate data to build (train) classifiers. In general, there is a variety of data preprocessing operations concerning both cases (instances) and features in datasets (cf. [1], [2], [3]). In [4], our consideration was focused on the case selection problem. Now, we deal with the feature selection problem.

Efficient analysis and retrieval of regularity from data is an extremely important task in the case of aggregation of vast amounts of data. Data mining processes are exposed to many aspects which cause failures. The large number of objects and variables, insignificance of some variables for the classification, interdependences between some part of variables, uneven

distribution of target classes, and other difficulties are the reason to develop methods for effective selection of significant feature subsets.

There are three major categories of feature selection methods: filter, wrapper and embedded methods. The first one scores variables individually using different measures and eliminates some of them before a model is constructed [5]. In turn, wrapper methods investigate the prediction accuracy of a model directly measuring the value of a feature set. Although effective, the exponential number of possible subsets places computational limits for the wide data sets that are the focus of this work. The last type, embedded methods firstly develop a learning model and then analyze the model to estimate the relevance of a feature. Effects are dependent on methods used for model generation. During our experiments the Boruta algorithm [6] for feature selection was used.

Experiments showed that selected datasets enabled us to build simpler and more accurate classifiers, both decision tree based and rule based ones.

II. INPUT DATA

Experiments which test the relative effectiveness of our approach have been performed on data describing over 70 cases (subjects) classified into three categories: high-functioning (*HIGH*), medium-functioning (*MEDIUM*), or low-functioning (*LOW*) autism. Each subject has been evaluated using an original sheet including questions about competencies grouped into 17 spheres marked with Roman numerals (only spheres used in our experiments are listed):

- VI. Support for active communication.
- VII. Active communication concerning objects, people, parts of the body.
- VIII. Imitation, the length and complexity of the utterance.
- IX. Needs, emotions, moods.
- X. Object communication (the level of specific symbols).
- XI. Symbolic communication.
- XII. Requests.
- XIII. Choices.
- XIV. Communication in a pair (with a contemporary, with an adult).
- XV. Social communication competences.
- XVI. Communication in a group and in social situations (in a team, at school, in the closest social environment).

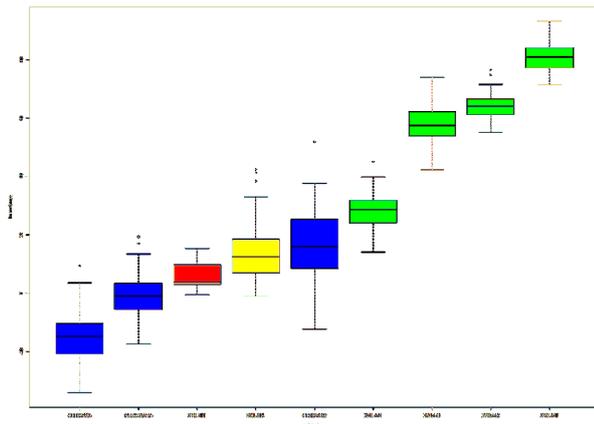


Fig. 2. Results of the feature selection process for sphere XVIII

TABLE I
A NUMBER OF FEATURES IN DATASETS

Dataset	#All features	#Confirmed features	#Tentative features
VI	18	5	4
VII	14	11	1
VIII	87	29	10
IX	51	21	6
X	3	1	0
XI	12	8	1
XII	9	1	2
XIII	14	11	3
XIV	13	11	1
XV	34	11	7
XVI	25	10	9
XVIII	6	4	1
XIX	7	6	1
XX	9	6	2
XXI	8	3	3
XXII	13	10	1
XXIII	13	8	2

- XVIII. Vocabulary.
- XIX. The degree of effectiveness of information.
- XX. The degree of motivation to communicate.
- XXI. The degree and type of hint in communication.
- XXII. Building the utterance - the degree of its complexity and functionality.
- XXIII. Dialogues.

Each case is described by over 300 features. Four values of features are possible, namely 0, 25, 50, and 100. They have the following meaning:

- 0 - not performed,
- 25 - performed after physical help,
- 50 - performed after verbal help/demonstration,
- 100 - performed unaided.

III. TOOLS

To solve a feature selection problem, we have used the Boruta algorithm. This algorithm applies random forest to determine all-relevant feature subset from datasets. It was designed as a wrapper method. Trees are independently developed on different bagging samples of the training set. The

importance estimation of an attribute is gathered as the loss of accuracy of classification caused by a random permutation of attribute values between objects. It is computed separately for all trees in the forest which use a given attribute for classification. After that, the average and standard deviation of the accuracy loss are computed. Thus, the Z score computed by dividing the average loss by its standard deviation can be used as an importance measure [6], [7]. Boruta separates attributes into three categories:

- confirmed,
- tentative,
- rejected.

Figures 1 and 2 show some examples of results of feature selection processes. The confirmed attributes are marked with green, tentative - with yellow, and rejected with red.

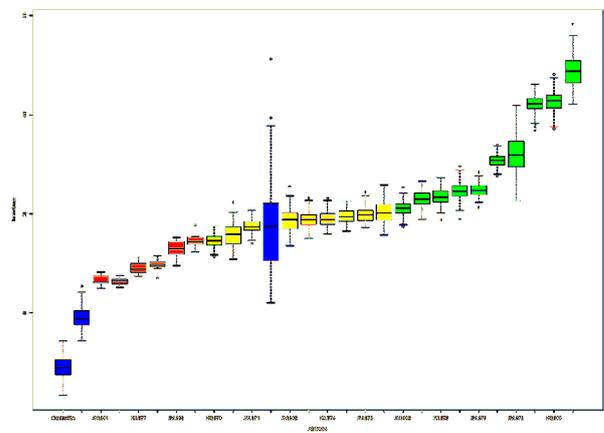


Fig. 1. Results of the feature selection process for sphere XVI

The datasets, after the feature selection processes, have been used to build decision tree and rule based classifiers.

For building classifiers, we have used two machine learning computer tools:

- RSES - a toolset for analyzing data with the use of methods coming from rough set theory [8].
- Orange - a comprehensive, component-based software suite for machine learning and data mining [9].

In RSES, we have used the LEM2 algorithm [10] for rule generation. LEM2 is most frequently used for rule induction. LEM2 explores the search space of feature-values pairs. It is based on lower and upper approximations of decision classes defined in rough set theory [11]. The expected degree of coverage of the training set by derived rules was set to 0.9. In a classification process, conflicts were resolved by standard voting (each rule has as many votes as supporting cases).

In Orange, we have used an algorithm for generation of decision trees based on the Gini criterion [1]. The following values of pruning parameters were set:

- minimum instances in leaves: 2,
- limit of the depth: 100.

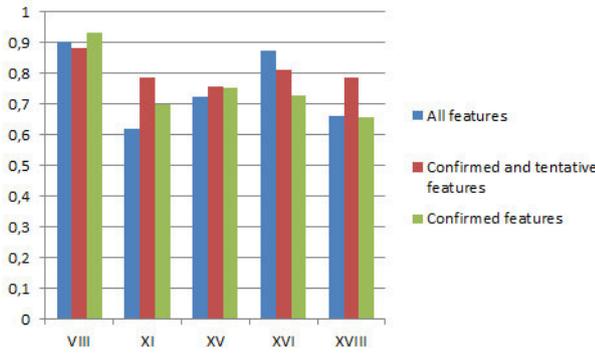


Fig. 5. Selected results of experiments with LEM2: classification accuracy

IV. RESULTS

In this section, we give selected results of experiments with the Boruta feature selection algorithm and classification algorithms (the algorithm of decision tree generation implemented in Orange and the LEM2 algorithm for rule generation implemented in RSES).

In our experiments, each data set has been treated separately. It enabled us to assess the evaluation sheet with respect to individual spheres. The results can be used in further development of the sheet. In the future, any adding, removing, and modifying of questions are allowed. Especially, the questions corresponding to rejected features should be checked.

Table I shows the effects of applying a feature selection procedure in terms of a number of features. Next, we present the results of assessment of classifiers for selected datasets (spheres), see Figures from 3 to 8. To estimate the accuracy of classifiers, ten-fold cross-validation method was used.

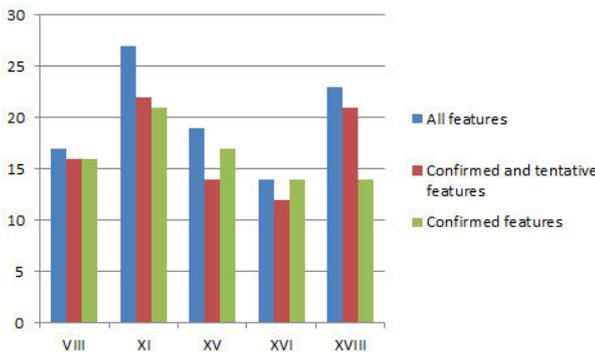


Fig. 3. Selected results of experiments with LEM2: a number of rules

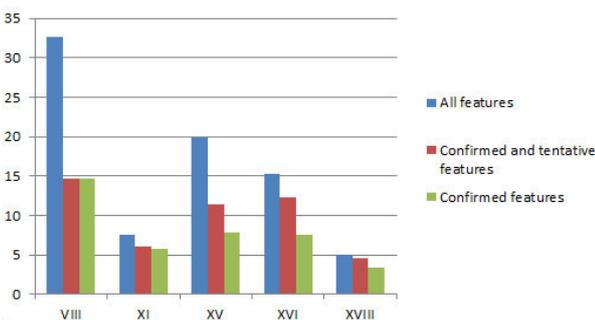


Fig. 4. Selected results of experiments with LEM2: mean of rule premise length

In case of complexity of classifiers, we have taken into consideration:

- a number of rules and mean of rule premise length (for a rule based classifier),
- a number of nodes and a number of leaves (for a decision tree based classifier).

In general, a feature selection procedure in the preprocessing step causes the decrease in the complexity of classifiers. In case of decision trees, a feature selection procedure positively influences the classification accuracy. In the case of rules generated by LEM2, taking into consideration the confirmed and tentative features seems to be more appropriate.

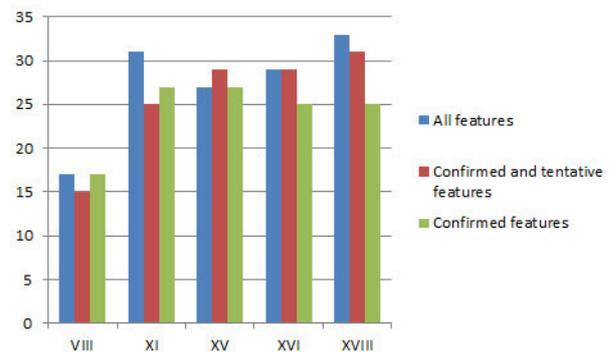


Fig. 6. Selected results of experiments with a decision tree: a number of nodes

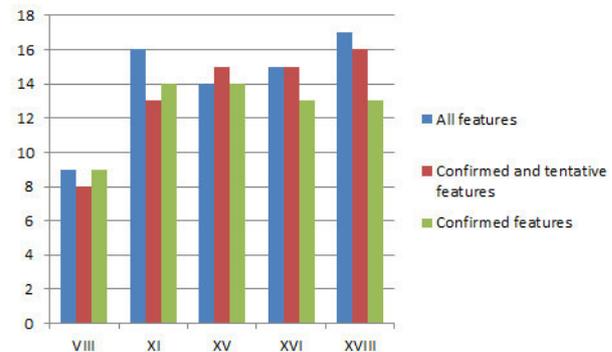


Fig. 7. Selected results of experiments with a decision tree: a number of leaves

V. CONCLUSIONS AND FURTHER WORK

In the paper, we have examined the Boruta algorithm to solve the feature selection problem for data coming from evaluation sheets of subjects with Autism Spectrum Disorders (ASDs). Simultaneously our research is also focused on the case selection problem [4]. Our main goal is to create hybrid

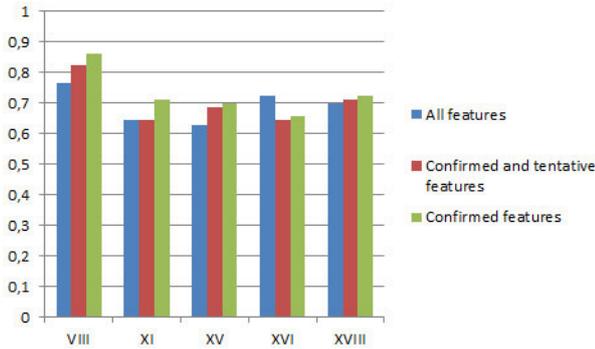


Fig. 8. Selected results of experiments with a decision tree: classification accuracy

classifiers combining a wide range of approaches that will be implemented in a dedicated computer tool supporting diagnosis and therapy of persons with ASDs.

REFERENCES

- [1] K. Cios, W. Pedrycz, R. Swiniarski, and L. Kurgan, *Data mining. A knowledge discovery approach*. New York: Springer, 2007.
- [2] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, ser. Intelligent Systems Reference Library. Switzerland: Springer International Publishing, 2015, vol. 72.
- [3] N. Jankowski and M. Grochowski, "Comparison of instances selection algorithms I. Algorithms survey," in *Artificial Intelligence and Soft Computing - ICAISC 2004*, ser. Lecture Notes in Computer Science, L. Rutkowski, J. H. Siekmann, R. Tadeusiewicz, and L. A. Zadeh, Eds. Berlin, Heidelberg: Springer-Verlag, 2004, vol. 3070, pp. 598–603.
- [4] K. Pancerz, A. Derkacz, and J. Gomuła, "Consistency-based preprocessing for classification of data coming from evaluation sheets of subjects with ASDs," in *Position Papers of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS'2015)*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 6, Lodz, Poland, 2015. doi: 10.15439/2015F393 pp. 63–67.
- [5] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, "Feature selection with ensembles, artificial variables, and redundancy elimination," *Journal of Machine Learning Research*, vol. 10, pp. 1341–1366, 2009.
- [6] W. R. Rudnicki, M. Wrzesień, and W. Paja, "All relevant feature selection methods and applications," in *Feature Selection for Data and Pattern Recognition*, ser. Studies in Computational Intelligence, U. Stańczyk and C. L. Jain, Eds. Berlin, Heidelberg: Springer-Verlag, 2015, vol. 584, pp. 11–28.
- [7] M. Kurasa and W. Rudnicki, "Feature selection with the Boruta package," *Journal of Statistical Software*, vol. 36, no. 1, 2010. doi: 10.18637/jss.v036.i11
- [8] J. G. Bazan and M. S. Szczuka, "The Rough Set Exploration System," in *Transactions on Rough Sets III*, ser. Lecture Notes in Artificial Intelligence, J. Peters and A. Skowron, Eds. Berlin Heidelberg: Springer-Verlag, 2005, vol. 3400, pp. 37–56.
- [9] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in Python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.
- [10] J. Grzymala-Busse, "A new version of the rule induction system LERS," *Fundamenta Informaticae*, vol. 31, pp. 27–39, 1997.
- [11] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, vol. 177, pp. 3–27, 2007. doi: 10.1016/j.ins.2006.06.003