computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

UNBIASEDNESS OF FEATURE SELECTION BY HYBRID FILTERING

Wiesław Pietruszkiewicz

West Pomeranian University of Technology
49, Żołnierska Str., 71-210 Szczecin, Poland,
wpietruszkiewicz@wi.zut.edu.pl

Abstract: *In this article we examine characteristics of feature selection algorithms by introducing their aspects important in practice. We will focus on the unbiasedness, analyse it and investigate a robust hybrid method of feature selection, being a composition of several feature filters, that could ensure unbiased results of selection. Using parallel multi-measures and voting, we reduce the risk of selecting non-optimal features, a common situation when we select attributes using single evaluation based on one evaluation criterion. To test this method we selected a personal bankruptcy dataset, containing various types of attributes and one of the popular machine learning benchmarks. By the performed experiments we will demonstrate that an approach of multi-evaluation used for features filtering may lead to the creation of effective and fast methods of features selection with an unbiased outcome.*

Keywords: *Feature selection, hybrid algorithms, machine learning, classification.*

1. INTRODUCTION

The most of AI applications assume that the deployment of particular features cannot be decided at the data gathering stage (apart the obvious irrelevant features) but have to be filtered out later. Hence, it is required to reduce the data dimensionality as this:

- Reduces the curse of dimensionality – the convergence of estimators used in the learning is much slower for problems with a higher dimensionality than for these with a lower number of dimensions.
- Lowers the memory requirements for data storage and processing – redundant or insignificant information increases the demand on memory, increasing storage costs or time span of stored data exceeding the possessed resources.
- Simplifies the model – which, being simpler, could be easily understandable by humans or software implementable.
- Speed-ups the process of learning – the complexity for machine learning methods usually is above linear complexity i.e. quadratic or cubic.
- Removes unnecessary attributes being a noise – irrelevant features could blur the problem and cause a lower quality of the results.
- Increases the generalisation ability – unnecessary attributes limit the model's generalisation ability i.e. the capability to work with the previously unseen data.

There are two solutions to the dimensionality reduction. It can be done by the recalculation of attributes into a smaller subset e.g. tasks done by Principal Components Analysis or Discriminant Analysis methods. The other approach to space dimensionality reduction is a feature selection, that generally returns a subset of attributes, being the most significant features for the modelled process.

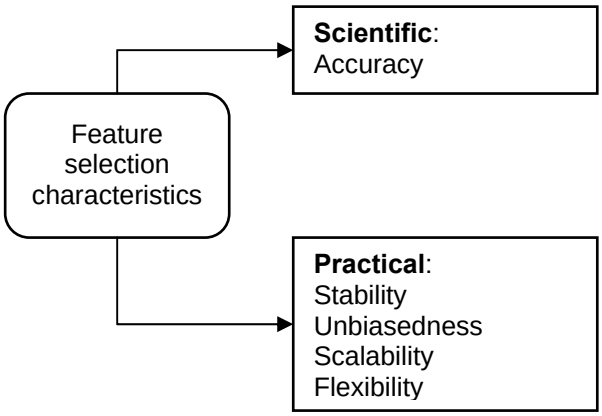

Fig. 1 – The characteristics of feature selection algorithms

Before analysing the types of feature selection algorithms we will briefly discuss how these algorithms are perceived from the scientific and practical perspective. While purely scientific perception is usually limited to the accuracy of selection (evaluated by other algorithms e.g. classifiers or regression models) their practical characteristic includes other features (see Figure 1)

very often omitted by researchers. In this article we will analyse *unbiasedness* – being the outcome of feature selection not biased i.e. that does not favour any algorithms that will be deployed later. Biased results of feature selection are caused by evaluation of features done by selected algorithms that does not necessarily leads to an optimal selected subset at the stage of application.

First of these features is important in situation when data storage is designed without the knowledge of particular algorithms that will be used later and due to the costs of data gathering or storage it is not possible to keep all features (redundancy of information or irrelevant attributes being a noise). Thus, experiments on a small subset of data might be used to select features and design production-ready data storage mechanisms.

researchers compared different filtering algorithms i.e. InfoGain, Chi^2 and correlation.

The second kind of feature selection algorithms – wrappers – was a subject of research presented e.g. [16] presented a sequential algorithm with low computational costs, being an example of general family of Forward Feature Selection algorithms. The other paper [10] proposed an algorithm of incremental feature selection.

The most of feature selection algorithms use batch processing, however [19] presented a streamwise algorithm allowing to dynamically select the new-coming features. The research in [12] has shown a semi-supervised feature selection.

The usage of filters or wrappers causes two major problems. For wrappers it is an extensive searching through the different combinations of attributes. It

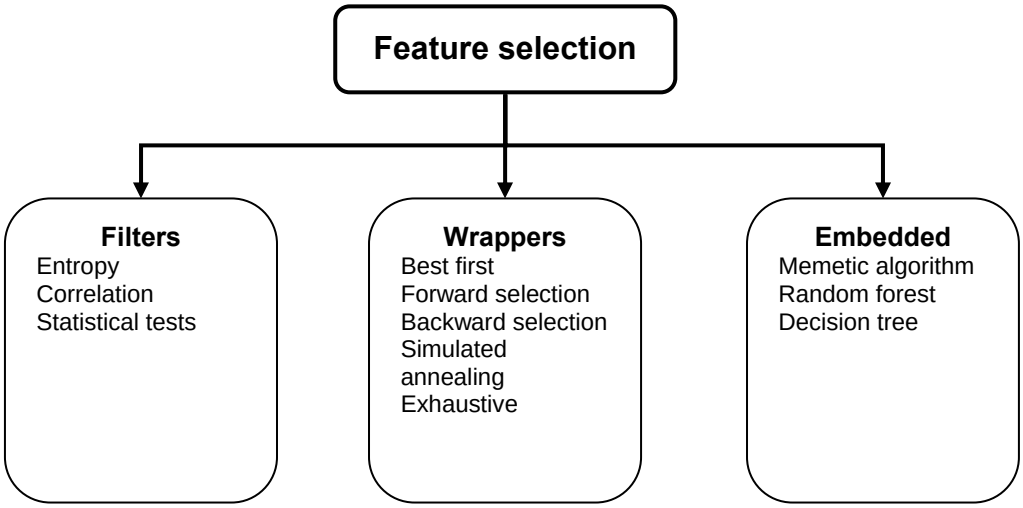

Fig. 2 – The division of feature selection algorithms

According to [6] feature selection algorithms may be divided onto two major groups: filters and wrappers. Filters use a measure to evaluate the relationship between the input and output attributes, while the wrappers are multi-step procedures testing different combinations of features. It is also possible to add another group of feature selectors i.e. embedded algorithms (see Figure 2), however they aren't used particularly for the feature selection but are incorporated by the others learning algorithms and deployed e.g. in pruning or node selection (for more information about all three kinds of feature selection algorithms see [15]).

The process of feature selection was a subject of many researches e.g. [7] compared correlation based filters with wrappers. Other papers focussing on filters are [8] presenting an algorithm based on Information Theory, similarly [2] compared different feature selection algorithms based on information entropy or [1] explained how both – feature and basis selection can be supported by a masking matrix. In another paper [18], the

forces the quality evaluation for each model build over each subset and every additional attribute increases the search space. On the other hand, the major problem we encounter using filters is an influence of measure selection on the quality of further developed model. The each evaluator used by filters, coming from information theory or statistics, may not be an optimal solution for various dimensions of subsets i.e. a subset filtered by one algorithm may be inferior to a potential subset selected by another algorithm.

In this paper we examine a hybrid approach to feature selection, done by a parallel multi-measure filtering (later called by Multi Measure Voting – MMV in abbrev.). In the following parts of article we present the results of experiments over multi-measure filtering used to select the features in a household bankruptcy prediction modelling.

2. UNBIASED HYBRID SELECTION

To select an optimal subsets of features we have selected the most popular algorithms of feature

Table 1. The names of attributes, type, description and kind

Attribute	Name	Type	Description	Group
X_1	Family members	numeric	Number of persons in household	demographical
X_2	Children	numeric	Number of children in household	demographical
X_3	Employed	numeric	Number of persons employed	demographical
X_4	Income	nominal	Total household net income	demographical
X_5	Gender	nominal	Respondent's gender	demographical
X_6	Women	numeric	Number of women in household	demographical
X_7	Men	numeric	Number of men in household	demographical
X_8	Average age	numeric	Average age of family	demographical
X_9	Responder's age	numeric	Responder's age	demographical
X_{10}	Education	numeric	Overall education of all household members	demographical
X_{11}	Domicile	nominal	Place of domicile	geographical
X_{12}	Marital status	nominal	Responder's marital status	demographical
X_{13}	Denomination	nominal	Responder's domination (All responders were Christians or atheists)	demographical
X_{14}	Handicapped	nominal	Is there any handicapped person in family?	demographical
X_{15}	Illness	nominal	Is there any member with a chronic illness?	demographical
X_{16}	Savings decisions	nominal	Who is responsible for taking the decisions about saving?	behavioural
X_{17}	Credit decisions	nominal	Who is responsible for taking the decisions about lending?	behavioural

selection i.e. InfoGain, GainRatio, Chi^2 and compared them together with MMV algorithm, incorporating all of them in multi-measure voting, proposed in this paper.

The first algorithm used to select the most important attributes was Info Gain, that in its core calculates the change of entropy from state X to $X|A$ (information gain caused by feature A). Assuming that $H(..)$ is an entropy function, the information gain may be calculated as $IG(X,A) = H(X) – H(X|A)$. The second algorithm was GainRatio, evaluating the significance of each attribute by measuring the gain ratio with respect to the class. We may represent this in form of: $GainR(Class, Attribute) = (H(Class) – H(Class|Attribute)) / H(Attribute)$. More information about the both algorithms – InfoGain and GainRatio may be found in [17]. The third method of feature selection was Chi^2, these algorithm evaluates the value of Chi^2 statistic with respect to the classes [5].

The fourth – a hybrid and robust algorithm tested herein was named MMV (Multi-Measurement and Voting). This algorithm is an analogy to a meta-classification algorithms that use a comity of parallel classifiers voting for a common decision [9]. By a parallel multiple filtering, that would use different measures, the risk of falling into the gap of non-optimality is reduced. In this paper we have used all three algorithms presented above to construct MMV, but its construction may vary and involve the other algorithms. The general schema of this method was presented on Figure 3.

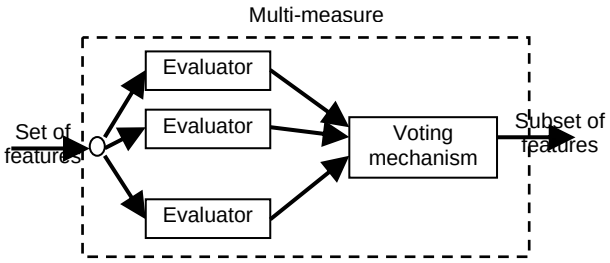

Fig. 3 – The division of feature selection algorithms

3. HOUSEHOLD BANKRUPTCY DATASET

In the research presented herein we have used a dataset about personal bankruptcy [14]. This dataset contained 17 input attributes and 1 output class attribute (see Table 1). The input attributes were 8 numeric and 9 nominal attributes. All these features were divided into three groups:

- Behavioural features – Describing how the financial decision are being taken by household.
- Demographical features – Describing the family i.e. the number of family members, their average age, education or family income.
- Geographical features – Containing the information about family's domicile.

The output attribute was a class feature, as all families were divided into three groups if family:

- Repaid or repay debts in advance or according to the schedule.

- Had or have slight problems in the repayments.
- Had or have significant problems in the repayments, stopped them or were a subject of any debt enforcement procedure.

This mixture of numeric and nominal attributes, having different characteristics, was selected as the testing data requiring feature selectors to prove their abilities to work with different types of data.

The results of evaluation for all algorithms were presented in Table 2 and as it can be noticed the ranks proposed by each algorithm differed. Therefore, the subsets filtered-out by each algorithm contained various attributes.

To evaluate the results of feature selection we have used a popular and flexible classification algorithms – C4.5 decision tree [11] and neural networks (in multi-layered perceptron variant [13]). During the experiments we have examined models with different numbers of attributes for each algorithm. To check their generalisation abilities we have tested accuracy for 3-fold Cross-Validation and Training Set. An objective environment for the comparison of models was ensured by keeping C4.5 parameters constant, otherwise adjustment could be done in favour of any algorithm. We set these C4.5 parameters to:

- The confidence factor was set to 0.25,
- The minimal support of leaf was set to 2,
- The number of data fold was set to 3 (one of folds was used in error pruning).

The neural networks were build and taught with these parameters:

- The learning ratio was set to 0.2,
- The momentum was set to 2,

The net had one hidden layer with a adaptive number of neurons.

The results of the experiments present the accuracy for experiments with 3-fold cross-validation or done on the training set Figures 4, 6 and Figures 5, 7 respectively. It is possible to observe that MMV (on charts denoted as Voting to emphasise its construction) was in most of situations as accurate as the best algorithm and it was very stable comparing to the other algorithms. It must be remembered that there was not any globally optimal algorithm, therefore Multi-Measure proved to be a fast and also effective approach to feature selection.

The Figures 8-11 represents the values of Receiver Operating Characteristic (more about ROC in [4]) for each class separately and the overall value for all classes. It is possible to observe that there exist an optimal ROC value for a subset of the attributes. Smaller subset of variables does not represent all useful information, while the larger dataset contains the attributes being a noise. These values were calculated for 3-fold Cross-Validation and it is possible to observe that similarly to the accuracy charts a hybrid approach to feature selection (MMV denoted in Figure 3 as Voting) was a very robust algorithm.

It must be pointed out that by analysing the primary evaluators we have observed that GainRatio was the most unstable method of features filtering as well as that the GainRatio-based models, sizing from 1 to 4, were highly inferior for to all the others algorithms.

Table 2. Features ranks for household dataset

Attribute	InfoGain	GainRatio	Chi^2	MMV
X_1	4	4	4	4
X_2	8	8	8	8
X_3	6	2	5	5
X_4	1	7	1	1
X_5	11	10	11	11
X_6	17	17	17	17
X_7	10	11	10	9
X_8	2	6	3	3
X_9	7	1	7	7
X_{10}	3	5	2	2
X_{11}	9	13	9	10
X_{12}	5	3	6	6
X_{13}	12	14	12	12
X_{14}	14	9	15	13
X_{15}	16	16	16	16
X_{16}	13	12	13	14
X_{17}	15	15	14	15

Table 3. Features ranks for image segmentation dataset

Attribute	InfoGain	GainRatio	Chi^2	MMV
X_1	16	17	16	16
X_2	9	9	10	9
X_3	19	19	19	19
X_4	18	18	18	18
X_5	17	16	17	17
X_6	13	13	13	13
X_7	15	15	15	15
X_8	12	12	12	12
X_9	14	14	14	14
X_{10}	2	5	2	2
X_{11}	1	2	1	1
X_{12}	6	3	6	6
X_{13}	3	6	4	4
X_{14}	11	11	11	11
X_{15}	10	10	9	10
X_{16}	7	8	7	7
X_{17}	4	4	5	5
X_{18}	8	7	8	8
X_{19}	5	1	3	3

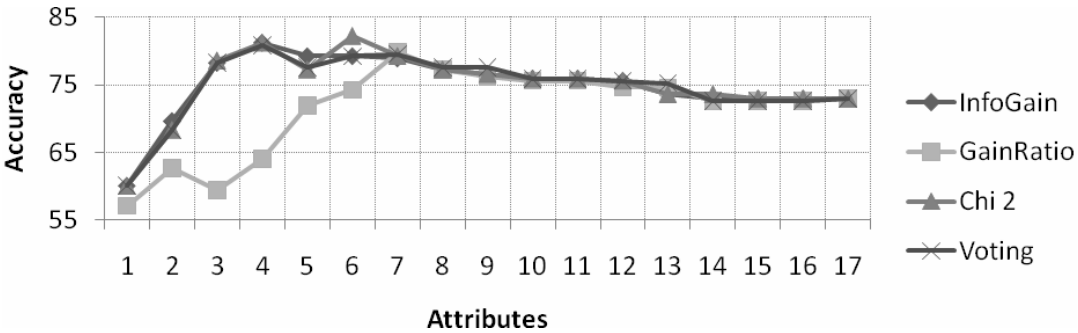

Fig. 4 – Accuracy for C4.5 (CV3, household bankruptcy experiment)

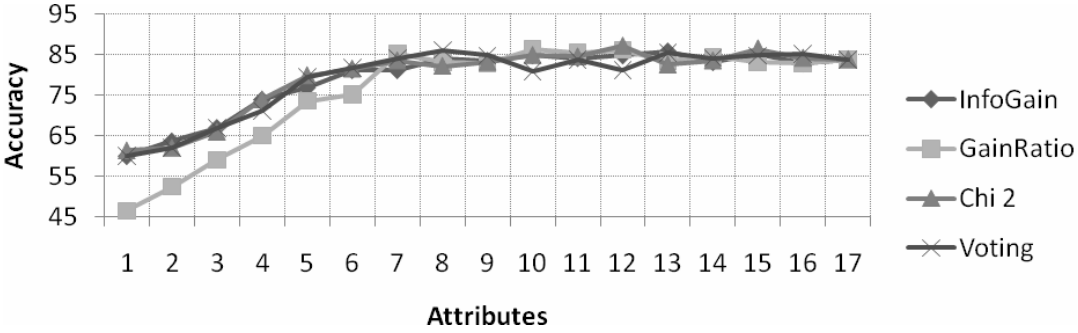

Fig. 5 – Accuracy for Neural Networks (CV3, household bankruptcy experiment)

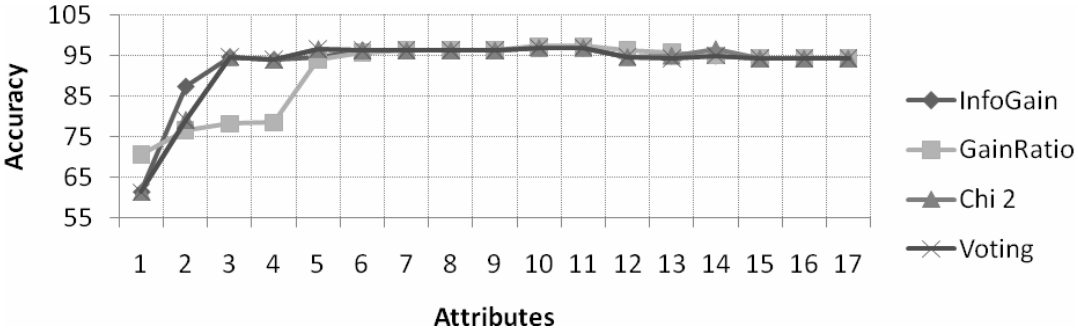

Fig. 6 – Accuracy for C4.5 (training set, household bankruptcy experiment)

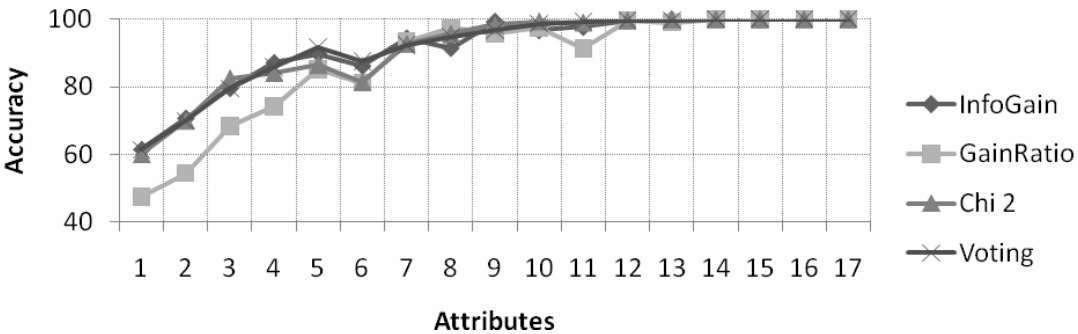

Fig. 7 – Accuracy for Neural Networks (training set, household bankruptcy experiment)

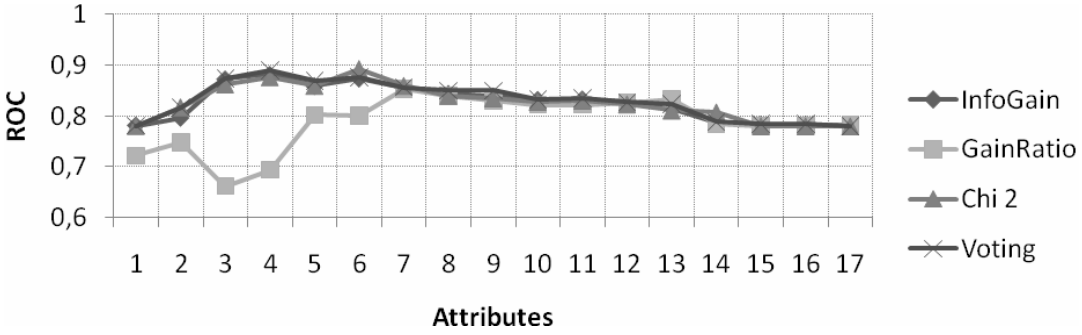

Fig. 8 – ROC for class 1 (household bankruptcy experiment)

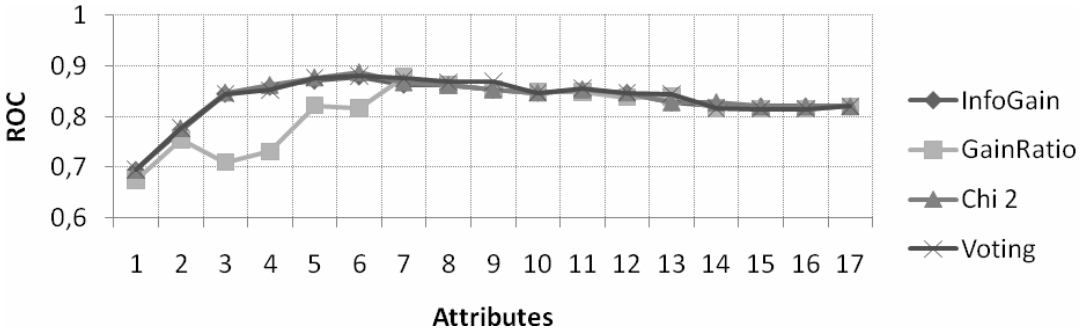

Fig. 9 – ROC for class 2 (household bankruptcy experiment)

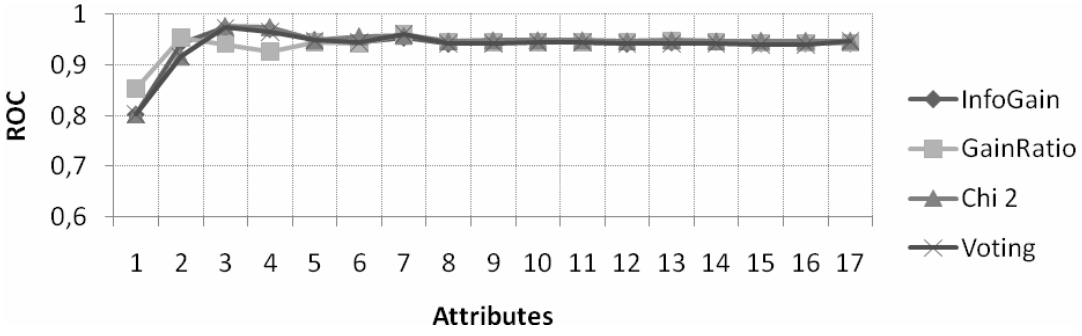

Fig. 10 – ROC for class 2 (household bankruptcy experiment)

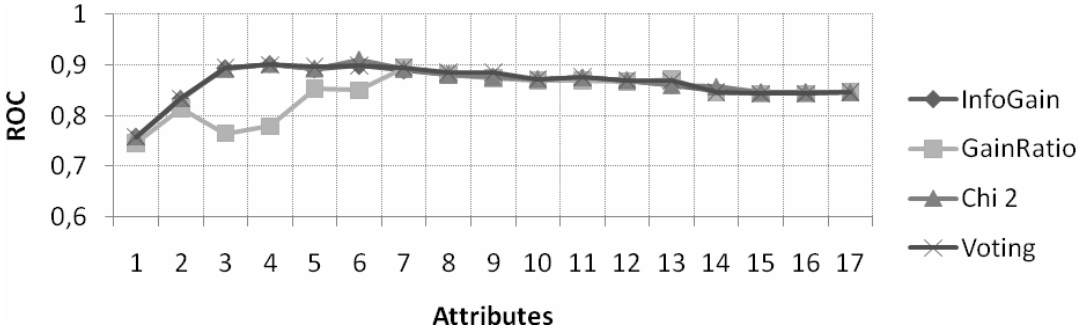

Fig. 11 – ROC for all classes (household bankruptcy experiment)

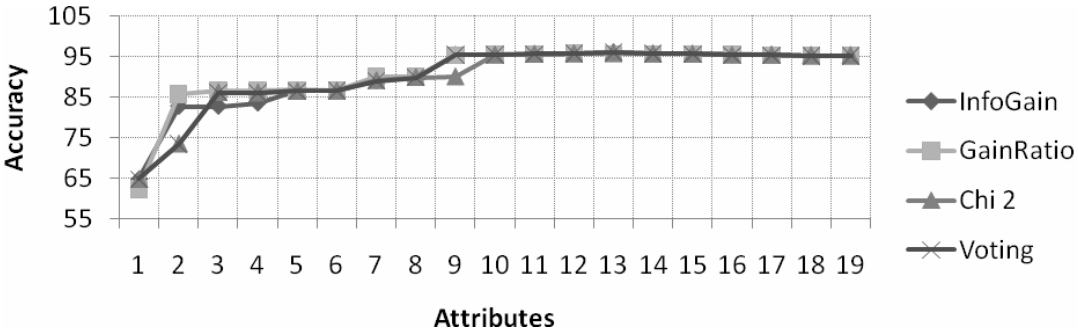

Fig. 12 – Accuracy for C4.5 (CV3, images segmentation set)

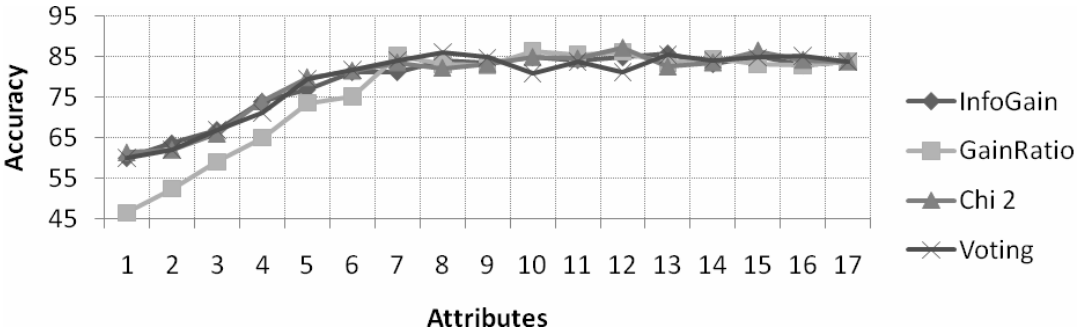

Fig. 13 – Accuracy for C4.5 (training set, images segmentation set)

4. IMAGE SEGMENTATION DATASET

To verify results presented in the previous section we examined results of classification for Statlog Image Segmentation dataset, being popular machine learning benchmark available at UCI dataset repository. This dataset contains 19 features with real values and 1 output class attribute. This dataset was selected due to different profile comparing to the household bankruptcy experiment, as the previous dataset contained mixed attributes, while this one had numeric real-valued features.

Results of selection for single algorithms as well as for the hybrid selection were presented in Table 3. As it may be noticed approximately first half features was ranked differently by algorithms, while the other half had identical ranks.

Figures 12 and 13 present the accuracy of classification done via C4.5 algorithm evaluated using training set or 3-fold cross validation. It must be noted that GainRation filter was the worst algorithm for subsets of attributes in range from 1 to7, while this algorithm was the best in overall for household experiment. Therefore, a hybrid selection should be considered as a robust and unbiased feature selection.

5. CONCLUSIONS

The research presented in this paper exploits practical characteristic of feature selection and analysed a hybrid selection, being a robust and unbiased method of the features filtering. The idea of this method assumes that it is possible to select features effectively and quickly by incorporating several basic methods of features evaluation. The voting done by all incorporated methods will allow this meta-evaluator to omit the risk of selecting low quality features due to biased filtering. Moreover, it can be done without the necessity of evaluation of different models, build using features recommended each of the primary feature evaluators – being a common solution to this problem.

We see several areas of future investigation. Firstly, we would like to extend the voting mechanism by incorporating the weighting. Secondly, we also plan to investigate other combinations of feature selection algorithms. Thirdly, we aim to rearrange the algorithm to involve a supervising mechanism deciding about the strength of signals generated by each singular filtering algorithm and how it should influence the overall filtering procedure.

The next stage of research will focus on the other elements of feature selection algorithms characteristics including stability, scalability and

flexibility. The analysis of all these areas, together with unbiasedness done in this paper, will make practical applications of feature selection easier and effective.

6. REFERENCES

[1] S. Avidan. Joint feature-basis subset selection. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[2] W. Duch, T. Wieczorek, J. Biesiada, M. Blachnik. Comparison of feature ranking methods based on information en-tropy. *Proceeding of International Joint Conference on Neural Networks*, 2004.

[3] A. Frank, A. Asuncion. *UCI Machine Learning Repository* http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[4] M. Gonen. *Analyzing Receiver Operating Characteristic Curves Using SAS*. SAS Press, 2007.

[5] P. E. Greenwood, M. S. Nikulin. *A Guide to Chi-Squared Testing*. John Wiley Sons, New York, 1996

[6] I. Guyon, A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research* (3) (2003), pp. 1157-1182.

[7] M. A. Hall, L. A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. *Proceedings of the Twelfth International FLAIRS Conference*, 1999.

[8] D. Koller, M. Sahami. Toward optimal feature selection. *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996.

[9] L. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-IEEE, Hoboken, 2004.

[10] H. Liu, R. Setiono. Incremental feature selection. *Applied Intelligence* (9) (1998), pp. 217-230.

[11] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann. San Francisco, 1993.

[12] J. Ren, Z. Qiu, W. Fan, H. Cheng, P. S. Yu. Forward semi-supervised feature selection. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2008.

[13] R. Rojas. *Neural Networks – A Systematic Introduction*. Springer-Verlag. Berlin, 1996.

[14] L. Rozenberg, W. Pietruszkiewicz. *The methodic of diagnosis and prognosis of household bankruptcy*. Difin. Warszawa, 2008.

[15] Y. Saeys, I. Inza, P. Larranaga, D. J. Wren. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23 (19) (2007).

[16] D. Ververidis. C. Kotropoulos. Sequential forward feature selection with low computational cost. *Proceedings of European Signal Processing Conference,* 2005.

[17] I. H. Witten. E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. San Francisco, 2005.

[18] Z. Zheng. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter archive* 6 (1) (2004), pp. 80-89.

[19] J. Zhou, D. P. Foster, R. A. Stine, L. H. Ungar. Streamwise feature selection. *Journal of Machine Learning Research* (7) (2006), pp. 1861-1885.

Wiesław Pietruszkiewicz, *received PhD degree in Computer Science from Szczecin University of Technology (currently West Pomeranian University of Technology) in Szczecin, Poland. He is Assistant Professor at Faculty of Computer Science and Information Technologies, West Pomeranian University of Technology in Szczecin, where teaches about the software design and programming of artificial intelligence & knowledge-based systems.*

His research interest relate to the applied data processing and intelligent systems.