The Hybrid Feature Selection for the Prediction of Household Bankruptcy

Wiesław Pietruszkiewicz

Faculty of Computer Science and Information Technology
West Pomeranian University of Technology in Szczecin
ul. Zolnierska 49, 71-210 Szczecin, Poland
wieslaw@pietruszkiewicz.com

Abstract

In this article we examine a robust hybrid method of feature selection that is a composition of several basic feature filters. By this meta–method, using parallel multi–measures and voting, we want to avoid falling into gaps of assumptions, characteristic to each feature selection filtering algorithm. Multi–evaluation of features and combining their ranks, reduce the risk of selecting non–optimal features, a common situation when we select attributes using single evaluation. The commonly used solution to this problem is an evaluation of multiple variants of feature selection sets, done by a few separated evaluators, but it increases the number of experiments and it finally costs and time of modelling. To test the method presented herein we selected a personal bankruptcy dataset, containing various types of attributes. By the performed experiments we demonstrate that an approach of multi–evaluation used for features filtering may lead to the creation of effective and fast methods of features selection.

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 later filtered out. Hence it is required to reduce the data dimensionality as it:

1. 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.

2. 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.

3. Simplifies the model – which, being simpler, could be easily understandable by humans or software implementable.

4. Speed–ups the process of learning – the complexity for machine learning methods usually is above linear complexity i.e. quadratic or cubic

5. Removes unnecessary attributes being a noise – irrelevant features could blur the problem and cause a lower quality of the results

6. 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.

According to (Guyon and Elisseeff 2003) 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 1), 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 (Saeys, Inza, Larra?aga, and Wren 2007)).

The process of feature selection was a subject of many researches e.g. (Hall and Smith 1999) compared correlation based filters with wrappers. Other papers focussing on filters are (Koller and Sahami 1996) presenting an algorithm based on Information Theory, similarly (Duch, Wieczorek, Biesiada, and Blachnik 2004) compared different feature selection algorithms based on information entropy or (Avidan 2004) explained how both – feature and basis selection can be supported by a masking matrix. In another paper (Zheng 2004), the 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. (Ververidis and Kotropoulos 2005) presented a sequential algorithm with low computational costs, being an example of general family of Forward Feature Selection algorithms. The other paper (Liu and Setiono 1998) proposed an algorithm of incremental feature selection.

The most of feature selection algorithms use batch processing, however (Zhou, Foster, Stine, and Ungar 2006) presented a streamwise algorithm allowing to dynamically select the new–coming features. The research in (Ren, Qiu, Fan, Cheng, and Yu 2008) 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 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 Dataset

In the research presented herein we have used a dataset about personal bankruptcy (Rozenberg and Pietruszkiewicz 2008). 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:

1. *Behavioural features* – Describing how the financial decision are being taken by household.

2. *Demographical features* – Describing the family i.e. the number of family members, their average age, education or family income.

3. *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:

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

2. Had or have slight problems in the repayments.

3. 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.

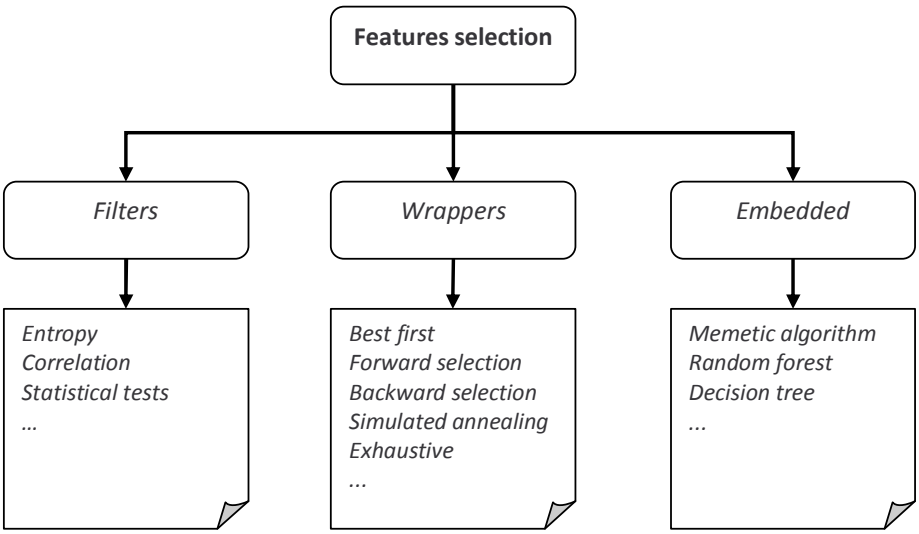


Figure 1: The division of feature selection algorithms

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

X_i	**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}	Martial status	nominal	Responder's martial status	demographical
X_{13}	Denomination	nominal	Responder's domination[1]	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

3 Algorithms of feature selection

To select an optimal subsets of features we have selected the most popular algorithms of feature 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 (Witten and Frank 2005). The third method of feature selection was Chi^2, these algorithm evaluates the value of Chi^2 statistic with respect to the classes (Greenwood and Nikulin 1996).

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 (Kuncheva 2004). 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 2.

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.

4 Results

To evaluate the results of feature selection we have used a popular and flexible classification algorithms – C4.5 decision tree (Quinlan 1993) and neural networks (in multi–layered perceptron variant (Rojas 1996)). 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).

while neural network was 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 (Figure 4) and (Figure 4 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 Figure 4 represents the values of Receiver Operating Characteristic (more about ROC in (Gonen 2007)) 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 Figures 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.

5 Summary

The research presented in this paper exploits the hybrid feature selection being a robust 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. 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

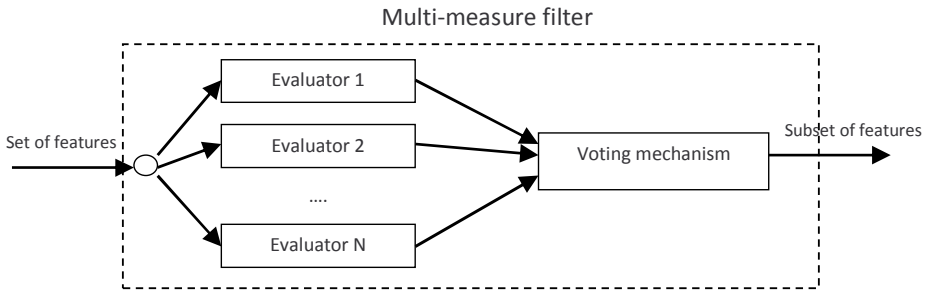


Figure 2: The schema of multi–measure feature filter

Table 2: Features ranks

X_i	**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

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.

References

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

Duch, W., T. Wieczorek, J. Biesiada, and M. Blachnik (2004). Comparison of feature ranking methods based on information entropy. In *Proceeding of International Joint Conference on Neural Networks*.

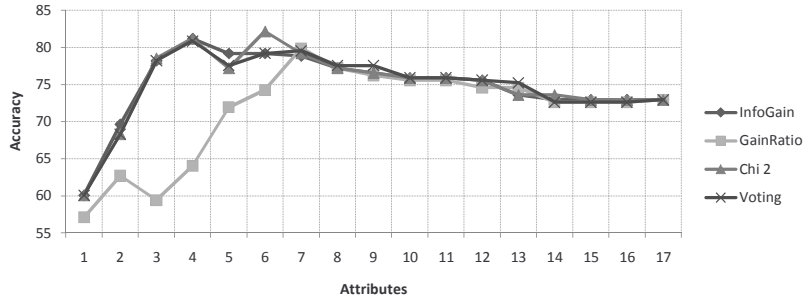
Gonen, M. (2007). *Analyzing Receiver Operating Characteristic Curves Using SAS*. SAS Press.

Greenwood, P. E. and M. S. Nikulin (1996). *A Guide to Chi-Squared Testing*. New York: John Wiley Sons.

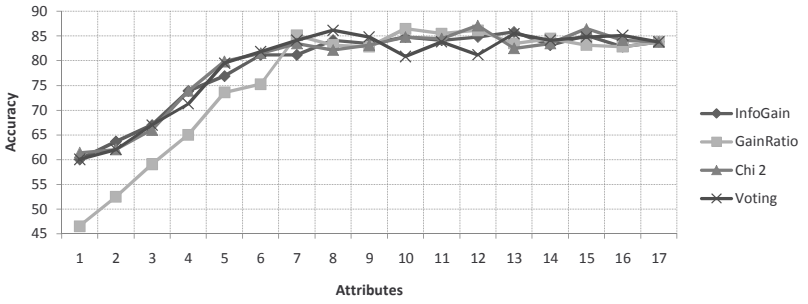
Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* (3), 1157–1182.

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

Koller, D. and M. Sahami (1996). Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*.

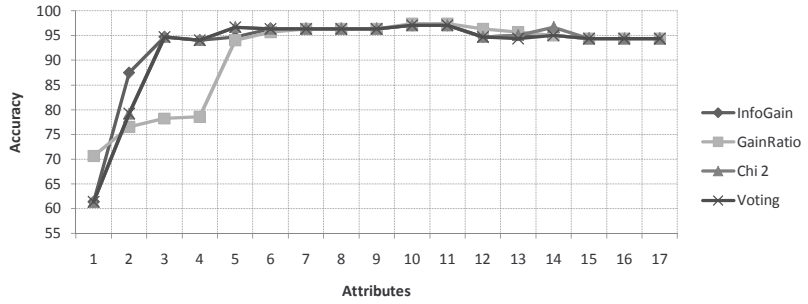


(a) Accuracy for C4.5 (CV3)

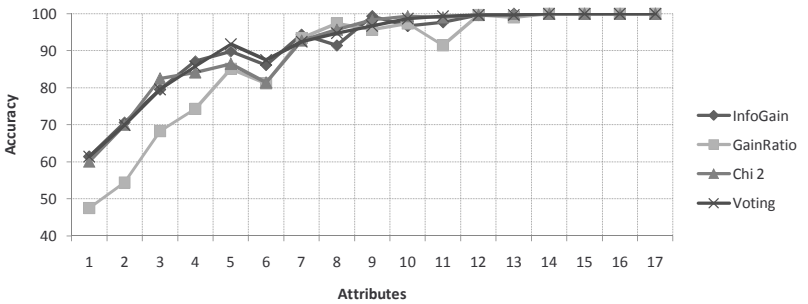


(b) Accuracy for Neural Networks (CV3)

Figure 3: The accuracy for C4.5 (a) and Neural Networks (b) evaluated by 3–fold cross–validation

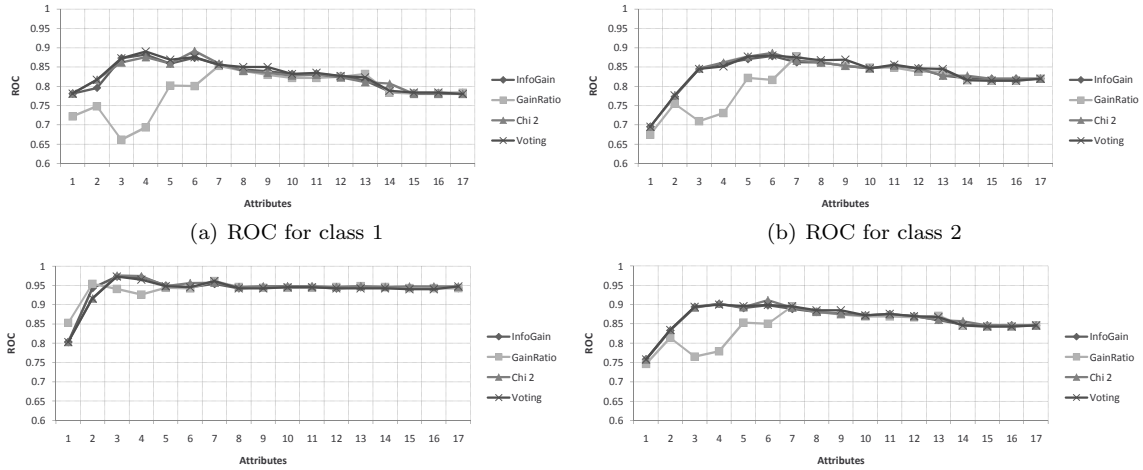


(a) Accuracy for C4.5 (training set)



(b) Accuracy for Neural Networks (training set)

Figure 4: The accuracy for C4.5 (a) and Neural Networks (b) evaluated on the training set



(a) ROC for class 1

(b) ROC for class 2

(c) ROC for class 3

(d) ROC for all classes

Figure 5: The values of ROC for class 1 (a), ROC for class 2 (b), ROC for class 3 (c) and ROC for all classes (d) vs number of features

Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken: Wiley-IEEE.

Liu, H. and R. Setiono (1998). Incremental feature selection. *Applied Intelligence 9*, 217–230.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco: Morgan Kaufmann.

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

Rojas, R. (1996). *Neural Networks – A Systematic Introduction*. Berlin: Springer-Verlag.

Rozenberg, L. and W. Pietruszkiewicz (2008). *The methodic of diagnosis and prognosis of household bankruptcy*. Szczecin: Difin.

Saeys, Y., I. Inza, P. Larra?aga, and D. J. Wren (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics 23*(19).

Ververidis, D. and C. Kotropoulos (2005). Sequential forward feature selection with low computational cost. In *Proceedings of European Signal Processing Conference*.

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

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

Zhou, J., D. P. Foster, R. A. Stine, and L. H. Ungar (2006, September). Streamwise feature selection. *Journal of Machine Learning Research 7*, 1861–1885.