ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 1(311), 2015

http://dx.doi.org/10.18778/0208-6018.311.03

Mariusz Kubus [*]

FEATURE SELECTION AND THE CHESSBOARD PROBLEM

Abstract. Feature selection methods are usually classified into three groups: filters, wrappers and embedded methods. The second important criterion of their classification is an individual or multivariate approach to evaluation of the feature relevance. The chessboard problem is an illustrative example, where two variables which have no individual influence on the dependent variable can be essential to separate the classes. The classifiers which deal well with such data structure are sensitive to irrelevant variables. The generalization error increases with the number of noisy variables. We discuss the feature selection methods in the context of chessboard-like structure in the data with numerous irrelevant variables.

Key words: chessboard problem, feature selection, feature relevance.

1. INTRODUCTION

Automatic feature selection has a key meaning in data mining when the goal is knowledge acquiring from a big datasets. There are a few areas where feature selection was successfully applied: i.e. text classification (Forman 2003), gene selection (Xing *et al.* 2001; Yu and Liu 2004), customer relationship management (Ng and Liu 2000). The variables which have no impact on dependent variable (called irrelevant or noisy) can lead to overfitting and a model has lower generalization ability. It means that the error on unseen data will be greater. In high dimensional spaces there is also the problem with the estimation of model parameters. The requirements to the number of observations in the training set grow exponentially with a grow of dimension, to save the accuracy of the estimation (*curse of dimensionality*). In this paper we put emphasis on the role of the search technique in feature selection methods. Considering the chessboard benchmark problem we discuss the term of feature relevance and we show the trade-off between feature subset selection and combinatorial complexity of this task.

[*] Ph.D., Department of Mathematics and Applied Computer Science, Opole University of Technology.

2. FEATURE SELECTION AS A SEARCH

Feature selection task can be formulated as a combinatorial optimization problem. Suppose we are given the set of multivariate observations with known values of response variable *Y* (training set):

$$U = \{(\boldsymbol{x_1}, y_1),...,(\boldsymbol{x_N}, y_N) : \boldsymbol{x_i} \in \boldsymbol{X} = (X_1,...,X_p), y_i \in Y, i \in \{1,...,N\}\} \ . \quad (1)$$

The response is nominal in discrimination and its categories are called classes. The goal is to learn the model with the lowest classification error on unseen data. Note that this is the most frequent model quality criterion but not the only possible. One can also introduce to the criterion the matrix of misclassification costs. We suspect that it is possible to obtain better model after projection of the data points on the subset $S \subset \boldsymbol{X}$. Even if the error is not significantly lower we are interested in using a lower number of predictors according to Occam's razor principle. To simplify the formalism we assume a family of models *F* and assume that the training set *U* is fixed. Let us choose some quality criterion *Q* of the feature subset *S*. Under our assumptions this criterion depends only on feature subset: $Q = Q(S)$. Thus, feature selection can be formulated as a problem of finding such a subset $S \in 2^{\boldsymbol{X}}$ so that the function $Q : 2^{\boldsymbol{X}} \to R$ reaches its optimum. Since the space of all feature subsets $2^{\boldsymbol{X}}$ is finite we are dealing with combinatorial optimization. In a high dimension, checking all possible subsets (exhaustive search) is impractical. It is especially problematic for computationally expensive learning algorithms like SVM. Moreover, exhaustive search can lead to overffiting (Quinlan and Cameron-Jones 1995; Jensen and Cohen 2000). Searching the $2^{\boldsymbol{X}}$ space is a hard combinatorial problem and usually heuristic techniques are implemented. It is of crucial importance in the methods of feature selection, what will be illustrated with the chessboard benchmark problem.

Due to the relationship between searching $2^{\boldsymbol{X}}$ space and the space of model parameters, feature selection methods were classified into three groups: filters, wrappers and embedded methods (Blum and Langley 1997). Filters work as a pre-processing step and the searches of a $2^{\boldsymbol{X}}$ space and model parameter space are performed independently. The criterion of the features quality *Q* is determined heuristically, thus it is not directly connected with a model quality. In fact, only wrapper approach uses model quality for evaluation of a feature subset. Searching the $2^{\boldsymbol{X}}$ space is performed as an outer loop of a learning algorithm. Embedded methods is a group of discrimination (or regression) methods where feature selection mechanism is built in a learning algorithm.

Here, searching a 2^X space and model parameter space are performed simultaneously.

In this article we would like to put emphasis on another classification of feature selection methods. As a feature quality criterion Q is a component of a search technique, we can ask if Q evaluates a single variable or rather a feature subset. Due to this question we divide feature selection methods into two groups: individual feature selection and feature subset selection. A related problem is defining the feature relevance. The priority goal is obtaining a model with the smallest classification error on unseen data, so the natural definition is that variable X is relevant $\Leftrightarrow X \in S_{best} : err(f(S_{best})) = \min$, where $f \in F$. Some authors call it usefulness, not relevance (i.e. Caruana and Freitag 1994). From the probabilistic point of view (Koller and Sahami 1996) X is relevant $\Leftrightarrow X \in S_{best} : P(Y|\mathbf{X}) = P(Y|S_{best})$, where conditional probabilities are equal or do not differ much. Note that these definitions silently assume exhaustive search. Applying heuristic or stochastic search we obtain subset $S_{optimal}$, and it is not guaranteed that $S_{optimal} = S_{best}$. Moreover, $S_{optimal}$ does not have to be the unique solution. In practice, we need the definition of feature relevance, which would be useful in the search process. In heuristic search usually one variable is added to (or removed from) the feature subset, and most proposed definitions (i.e. John *et al.* 1994; Blum and Langley 1997; Guyon and Elisseeff 2006) concern the individual relevance of the variable. Guyon and Elisseeff (2006) definied (approximately) sufficient feature subset using Kullback-Leibler divergence as (Koller and Sahami 1996). Multivariate approach allows to discover the interactions between variables, on the other hand one must face the combinatorial complexity.

3. THE CHESSBOARD BENCHMARK PROBLEM

The chessboard problem is the classical example where two, individually irrelevant variables, are important in discrimination task. Let us consider two classes discrimination problem in the plane OX_1X_2. Data points from the first class lie in the region $[0,1]\times[0,1]$ or in $[1,2]\times[1,2]$. Data points from the second class lie in the region $[1,2]\times[0,1]$ or in $[0,1]\times[1,2]$. The realizations of the variables X_1, X_2 are generated from the uniform distribution and every quadrat contains the same number of data points[1] (Fig. 1). The classes are clearly separable, and projection on any axis leads to 100% overlapping.

[1] An analogous example can be formulated for binary variables and it is known as XOR problem.

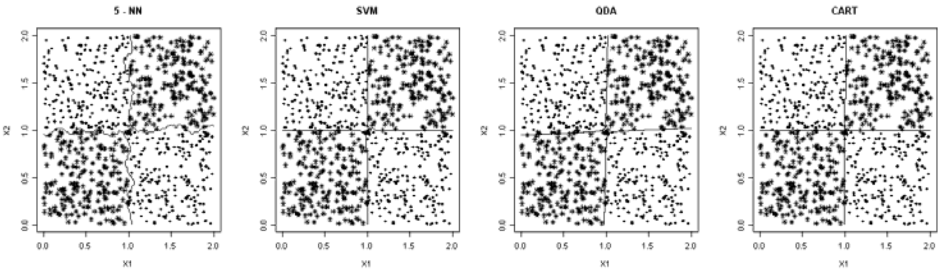

Fig. 1. The chessboard problem and boundaries between classes for 4 discrimination methods
Source: own computations.

It is no use applying the linear model in such a situation. Anyway, the non-parametric discrimination methods and quadratic discriminant analysis (QDA) can deal in such conditions quite well. Figure 1 depicts decision boundaries between classes for a few discrimination methods. The problem is that these methods are sensitive on the variables, which in no way affect the response variable. Let us add three individually irrelevant variables at a time in the following 20 iterations. Let the first two variables be generated – independently of classes – from $N(0;1)$, and the third be their linear combinations with a Gaussian noise. Figure 2 depicts classification error estimated in every iteration by splitting the data 30 times on training and test samples. Note, that adding even three irrelevant variables can dramatically increase the error as in 5-NN method. Thus, the problem arises how to discover the chessboard-like structure in the datasets with many irrelevant variables. In this context, we discus shortly the theoretical properties of feature selection methods from three groups: filters, wrappers and embedded methods.

The representatives of the third group (regularized versions of logistic regression or tree based models) do not suit this structure or do not work as a feature selectors (see Fig. 2).

Univariate scoring of the variable importance does not work in this case. The distributions of variables X_1, X_2 are the same in the classes, and the evaluation of X_1, X_2 will not differ from the evaluation of artificially included irrelevant variables. Therefore, it is necessary to apply the multivariate criterion of variable relevance. There are a few propositions in the literature, i.e. group correlation (Hall 2000), or Hellwig's criterion (Hellwig 1969) applied for instance with symmetrical uncertainty measure by Gatnar (2005). The problem is that these criteria are constructed as aggregation of an individual impact on the response variable, which also does not work in the chessboard case. The only multivariate filter which seems to be promising is Relief algorithm proposed by

Kira and Rendell (1992). Relief – inspired by nearest neighbour classifier – assigns the weights to the features in iterative procedure. The number of iterations is pre-specified and initial weights are equal to zero. In every iteration an example is sampled from the training set and its nearest neighbours are found: from the same class (the so- called nearest hit) and from the different class (nearest miss). The weight of each variable X_j is updated according to the formula:

$$W_j \leftarrow W_j - (x_j - x_j^{(h)})^2 + (x_j - x_j^{(m)})^2 . \tag{2}$$

Features with the weights less than zero are considered as irrelevant. More radical threshold can also be fixed. Kononenko (1994) proposed to take *k* nearest hits and *k* nearest misses.

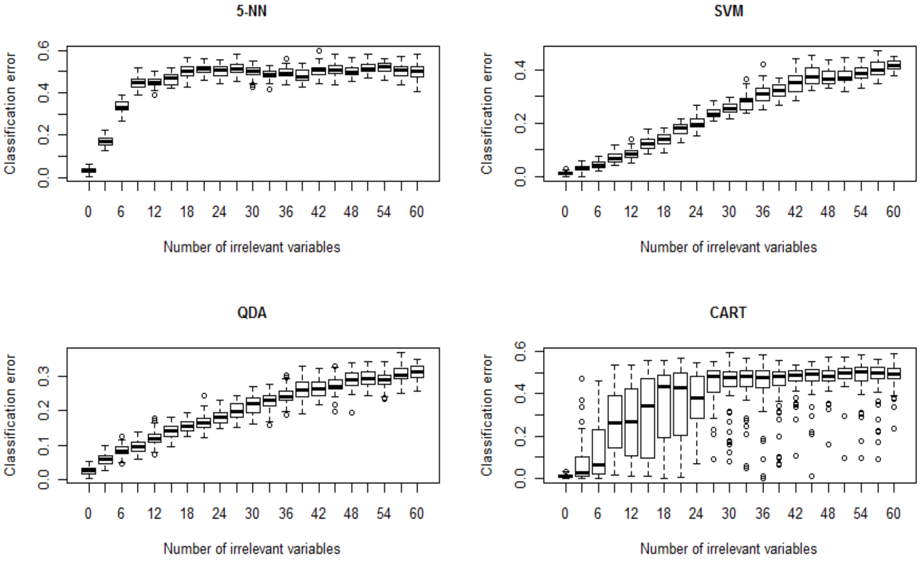

Fig. 2. Classification errors after adding noisy variables to the chessboard data
Source: own computations.

Wrappers seem to be perfect to discover the chessboard structure. The model quality criterion is also the assessment of a feature subset. Thus, learning the model using subset $S = \{X_1, X_2\}$ we should obtain significantly better evaluation than using other subsets. Performing exhaustive search we are sure to find the relevant variables. However, we would like to find a feature selector which could be applied in high dimensional spaces, where checking all feature subsets is impractical. Therefore, the search strategy has a key meaning and we should focus on it. The first issue is how to perform the steps between feature subsets.The common solution in heuristic search is adding to (or removing from) the current subset one variable at a time. In this case, forward selection will not work because of low evaluation of the true variables in the first stage. Backward elimination seems reasonable but this solution is extremely time consuming in the case of computationally expensive learning algorithms. For this reason Guyon *et al.* (2002) proposed recursive feature elimination (RFE) combined with SVM. It uses ranking of variables which is obtained from a model. RFE starts from a full set of variables. The model is learned and it should give ranking of variable importance. Then we remove the worst variable from this ranking, instead of considering all possible subsets with one variable discarded. It is repeated iteratively. In the next section we verify the two most promising feature selection methods (Relief and RFE-SVM) using artificially generated dataset.

4. SIMULATION STUDY

The chessboard-like structure was generated in 3-dimensional space. There were 4 clusters in every of two classes and all clusters were well separable from each other. Each cluster contained 100 examples from the spherical Gaussian distribution and centres were placed in the vertices of the cube, so that classes perfectly overlapped after projection on any axis or any plane spanned by the axes. In this way, variables X_1, X_2, X_3 are individually and pairwise irrelevant. Splitting the data 30 times into train and test sets we have obtained the following classification errors (in %) with standard errors: 3.66 (0.16) for 5-NN classifier, 4.04 (0.19) for SVM with radial kernel, and 53.97 (0.98) for QDA. The topology of the data cloud in 3-dimensional space makes the QDA method useless.

At the next stage of the experiment we introduced to this data 40 noisy variables with equal distributions in the classes. They represented various distributions:

- 15 variables from the normal distribution (every third was a sum of the previous two with a Gaussian noise added),
- 5 variables from the exponential distribution (lambda = 1:5),
- 5 variables from the mixture of $N(0,1)$ and $N(5,0.1*j)$ for j = 1:5 (1/3 observations were from standardised normal distribution),
- 5 variables from Bernoulli distribution (equal fractions of 0 and 1),
- 5 variables from Bernoulli distribution (fraction of 1 equal to 10%),
- 5 variables from Bernoulli distribution (fraction of 1 equal to 5%).

RFE-SVM was run 30 times with the split into training and test sets. Every time the true variables X$_1$, X_2, X_3 were discovered and all irrelevant variables were removed. Relief did not work so well (Fig. 3). We ran it 10 times using randomly selected subset of observations (half the training set). It can be seen that the variable X_3 is evaluated worse than some irrelevant variables. After carrying out more research we turned out, that it was caused by too many variables. Relief worked quite well for a smaller number of variables independently of their distributions (Fig. 3).

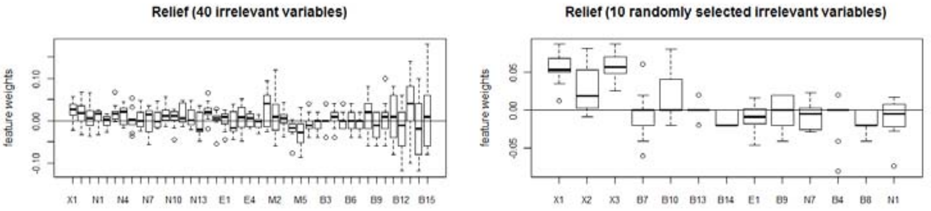

Fig. 3. Feature weights in Relief algorithm in 10 runs

Source: own computations.

As a single class in our artificial dataset is not homogenous, the question is, if cluster analysis can help to obtain better classifier. We applied *k*-means method to discover the clusters in two classes separately. Then, after dummy variables which indicate the clusters were introduced to the dataset, SVM was run again. The model yielded lower classification error: 3.42 (0.16). Note however, that we have obtained this result in the space X_1, X_2, X_3, thus without irrelevant variables. Adding even one irrelevant variable affected the sharp decrease of the silhouette index. Obtained the highest value of the silhouette index was less than 0.4 in the classification on 9 clusters, whereas the real number of clusters was the worst evaluated (silhouette index around 0.23). Thus, cluster analysis can improve the model quality, but noisy variables should be previously removed.

5. SUMMARY

Finding the best feature subset (for discrimination task) depends in practice on search technique. The interaction of the variables (relevance in the context) can be captured only using multivariate evaluation functions, but it encounters the problem of combinatorial complexity. Having carried out the simulation research on the dataset with chessboard-like structure we conclude that RFE-SVM is the best feature selector. We can also recommend Relief algorithm but the number of noisy variables in the dataset should not be too large. Applying cluster analysis as a pre-processing step can improve the quality of a classifier in the case of chessboard-like structure. However, it should be noted that noisy variables have a strong influence on the results of clustering.That is why such variables should be previously removed.

REFERENCES

Blum A.L., Langley P. (1997), Selection of relevant features and examples in machine learning, *Artificial Intelligence*, v. 97 n. 1–2, p. 245–271.

Caruana R.A., Freitag D. (1994), How useful is relevance? *Working Notes of the AAAI Fall Symposium on Relevance* (pp. 25–29). New Orleans, LA: AAAI Press.

Forman G. (2003), An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3: 1289–1305.

Gatnar E. (2005), Dobór zmiennych do zagregowanych modeli dyskryminacyjnych, in: Jajuga K., Walesiak M. (Eds.), *Taksonomia 12, Klasyfikacja i analiza danych – teoria i zastosowania,* Prace Naukowe Akademii Ekonomicznej we Wrocławiu, n. 1076, p.79–85.

Guyon I., Elisseeff A. (2006), An introduction to feature extraction, in I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), *Feature Extraction: Foundations and Applications,* Springer, New York.

Guyon I., Weston J., Barnhill S., Vapnik V. (2002), Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, 46: 389–422.

Hall M. (2000), Correlation-based feature selection for discrete and numeric class machine learning, *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco.

Hellwig Z. (1969), Problem optymalnego wyboru predykant, *„Przegląd Statystyczny”*, n. 3–4.

Jensen D. D., Cohen P. R. (2000), Multiple comparisons in induction algorithms. *Machine Learning*, 38(3): p. 309–338.

John G.H., Kohavi R., Pfleger P. (1994), Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, Morgan Kaufmann, p. 121–129.

Kira K., Rendell L. A. (1992), The feature selection problem: Traditional methods and a new algorithm. In *Proc. AAAI-92*, p. 129–134. MIT Press.

Koller D., Sahami M. (1996), Toward optimal feature selection. In *13th International Conference on Machine Learning*, p. 284–292.

Kononenko I. (1994), Estimating attributes: Analysis and extensions of RELIEF, In *Proceedings European Conference on Machine Learning*, p. 171–182.

Ng K. S., Liu H. (2000), Customer retention via data mining. *AI Review*, 14(6): 569 – 590.

Quinlan J.R., Cameron-Jones R.M. (1995), Oversearching and layered search in empirical learning. In Mellish C. (ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Morgan Kaufman, p. 1019–1024.

Xing E., Jordan M., Karp R. (2001), Feature selection for high-dimensional genomic microarray data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 601–608.

Yu L., Liu H. (2004), Redundancy based feature selection for microarray data. In *Proceedings of the Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 737–742.

Mariusz Kubus

SELEKCJA ZMIENNYCH A PROBLEM SZACHOWNICY

Streszczenie. W artykule podjęto dyskusję nad aspektem przeszukiwania w metodach selekcji zmiennych. Posłużono się znanym z literatury przykładem szachownicy, gdzie zmienne, które indywidualnie nie mają mocy dyskryminacyjnej (mają jednakowe rozkłady w klasach) mogą rozpinać przestrzeń, w której klasy są dobrze separowalne. Uogólniając ten przykład wygenerowano zbiór z trójwymiarową strukturą szachownicy i zmiennymi zakłócającymi, a następnie zweryfikowano metody selekcji zmiennych. Rozważono też możliwość zastosowania analizy skupień jako narzędzia wspomagającego etap dyskryminacji.

Słowa kluczowe: problem szachownicy, selekcja zmiennych, ważność zmiennych.