Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
6(339) 2018

DOI: http://dx.doi.org/10.18778/0208-6018.339.01

Mariusz Kubus
Opole University of Technology, Faculty of Production Engineering and Logistics,
Department of Mathematics and IT Applications, m.kubus@po.opole.pl

The Problem of Redundant Variables in Random Forests

Abstract: Random forests are currently one of the most preferable methods of supervised learning among practitioners. Their popularity is influenced by the possibility of applying this method without a time consuming pre-processing step. Random forests can be used for mixed types of features, irrespectively of their distributions. The method is robust to outliers, and feature selection is built into the learning algorithm. However, a decrease of classification accuracy can be observed in the presence of redundant variables. In this paper, we discuss two approaches to the problem of redundant variables. We consider two strategies of searching for best feature subset as well as two formulas of aggregating the features in the clusters. In the empirical experiment, we generate collinear predictors and include them in the real datasets. Dimensionality reduction methods usually improve the accuracy of random forests, but none of them clearly outperforms the others.

Keywords: random forests, redundant variables, feature selection, clustering of features

JEL: C1, C38, C52

1. Introduction

The objective of data mining techniques is the extraction of useful information from large datasets. In many applications, researchers process data containing many uninformative variables, outliers or missing values. In marketing research, for example, datasets consisting of variables of mixed types are a common occurrence. Therefore, practitioners usually prefer the methods which cope with all these problems. Additionally, the algorithms are expected to be fast. Analysts agree that cleaning data is more time consuming than the modelling stage. Classification trees (Gatnar, 2001) are a method that does not require a pre-processing step and can deal with all the listed problems. Note that this method belongs to embedded methods of feature selection (Guyon et al., 2006). Classification trees introduce into the model only variables which locally optimise a homogeneity criterion, and in this way they perform automatic feature selection. Unfortunately, these models are not very stable. Small changes in a training sample affect to a great extent the form of classifier. Moreover, the shape of decision boundaries results sometimes in lower accuracy of the trees. These drawbacks may be overcome by combining many single trees into one aggregated model. The last two decades have shown the rapid development of this approach which is called ensemble methods. The error of the ensemble is lower than the mean error of the base models (Breiman, 1996). Breiman (1996) and then Freund and Schapire (1996) have proved the improvement of stability. Moreover, a boosting method gives a possibility of bias reduction (Freund, Schapire, 1996). Ensemble methods work effectively if the base models have sufficient accuracy (at least slightly higher than prior probabilities) and they are diverse, that is, they do not classify identically the same observations from the training set. Thus, the training samples are usually bootstrapped.

One of the most popular ensemble methods is random forests (Breiman, 2001). This method can be viewed as a modification of the bootstrap aggregation (Breiman, 1996). The novel idea is to examine only small subsets of features which are randomly picked in each node. This approach supports diversity as well as decreases the computational cost. Random forests are sometimes called the best "off-the-shelf" classifiers. This term is used to encompass the methods which do not require a time consuming pre-processing step or careful tuning of the learning parameters. For this reason, random forests have found application in many areas of human activity, e.g.: churn analysis, fraud detection, prediction of bankruptcy, gene selection, and the supporting of medical diagnosis.

We have found interesting that although random forests have a built-in feature selection mechanism, many studies focus on the improvement of their performance via additional reduction of dimensionality. Granitto et al. (2006) applied the RFE procedure which utilises feature ranks to accelerate the search of the best feature subset. Gregorutti, Michel and Saint-Pierre (2017) used this algorithm with the

permutation importance measure considering high-dimensional tasks and correlated predictors. Toloşi and Lengauer (2011) considered medical data of such characteristics but they used clustering of features as the pre-processing step. Kursa and Rudnicki (2010) compared the relevance of the features to randomly generated features. Ye et al. (2013) proposed a stratified sampling method to select feature subspaces, while Hapfelmeier and Ulm (2013) selected features with the use of permutation tests.

In this paper, we attempt to examine the weaker side of random forests. We have found that the accuracy of this method decreases in a presence of redundant variables. Our goal is to verify the commonly proposed in that case feature selection methods as well as to examine the less popular approach which uses clustering of features. Our work differs from the above-mentioned (Toloşi, Lengauer, 2011) in terms of two issues. Firstly, we use the correlation-based dissimilarity measure rather than the Euclidian distance. Secondly, we consider datasets with a moderate number of features, which is more characteristic in economic sciences. The rest of this paper is organised as follows. In Section 2, we show the drawback of random forests. In Section 3, we briefly present methods which cope with redundant variables. Next, in Section 4, we report the experimental results, and finally Section 5 presents the summary of our work.

2. The drawback of random forests

Consider a classification task where the training set is given as:

$$U = \left\{(\boldsymbol{x}_1, y_1),...,(\boldsymbol{x}_N, y_N) : \boldsymbol{x}_i \in \boldsymbol{X} = (X_1,...,X_p), y_i \in Y, i \in \{1,...,N\}\right\}, \qquad (1)$$

and the objective is to estimate the model $y = f(\boldsymbol{x})$ which would predict a class for the objects that have not been observed yet. In ensemble learning, the set of base classifiers (f_1, …, f_M) is constructed, and M outputs are combined for the final prediction. Random forests (Breiman, 2001) use for this purpose simple majority voting where the most frequent class label among $f_m(\boldsymbol{x})$ is assigned to the object $\boldsymbol{x}$. The algorithm works as follows. A single binary tree is constructed without pruning in each iteration. The architecture is parallel, which means that each tree is built independently of the previous one. To obtain diversity of the base classifiers, the training sets are bootstrapped. The innovation in the comparison with the bagging method is randomisation of the feature subsets which are examined in the nodes of the trees. The recursive partitioning process is modified in this way that only q of p variables are considered in each node to select the best split-point. This characteristic of the algorithm also determines its speed. Random forests method scales well in high dimensional problems, e.g.: text classification. The number

of iterations M and the number of sampled features q are the input parameters. Many experiments which are reported in the literature point that the misclassification rate decreases with the number of trees, and stabilises after 100 or 200 iterations (see e.g.: Hastie, Tibshirani, Friedman, 2009). In turn, the recommended q for classification is equal to the square root of the number of input variables p (Breiman, 2001).

An additional advantage of random forests is the ranking of feature importance. At each split, the improvement of homogeneity criterion is used as the measure of feature importance. It is accumulated over all the nodes of all the trees. A more advanced method of feature evaluation uses out-of-bag and permuted values of the variables (Breiman, 2001). Both versions are available in the package `randomForest` of R program which we use in our experiment.

As every classifier, random forests work worse in certain conditions. One can suspect that if there is a great number of uninformative variables in relation to informative variables, then the probability that only irrelevant variables will be picked in some nodes is greater. In this case, irrelevant variables deteriorate the accuracy of the model. We have found that this is not the only problem with the variables. Assume that there are relevant but correlated predictors in the training set. Our intuition is that in deeper nodes, where the number of objects is lower, the learning algorithm may introduce some of such variables which are less correlated with the response Y. Thus, we suspect that informative but correlated to each other predictors can deteriorate the predictive abilities of random forests. Let us look at the following experiment. We have generated 300 observations from two quite well separated classes. Both classes are from five dimensional Gaussian distributions with means vectors (0, 0, 0, 0, 0) and (2, 2, 2, 2, 2) respectively, and unit covariance matrices. Moreover, we have included five irrelevant variables from $N(0; 1)$ as well as redundant variables which have been constructed as follows: for each relevant variable X_j, construct five correlated variables Q_{jk} according to the formula $Q_{jk} = X_j + Z_{jk}$, where $Z_{jk} - N(0; 0.1 \cdot \mathrm{sd}(X_j))$. We have run random forests twice, using only a subset of relevant and irrelevant variables, and using all artificial datasets. Classification errors have been estimated as an average test error from 30 splits. We have obtained misclassification rates 1.5% (with 0.2% standard error) in the first case and 2.2% (0.2%) in the second one. Assuming the significance level of 0.05, the difference is significant. The Wilcoxon rank sum test gives p-value equal to 0.01074. We have observed that random forests perfectly recognise irrelevant variables. This example shows that redundant variables can be more problematic for these classifiers than irrelevant ones.

3. Redundant variables and solutions of this problem

There are two main approaches to solve the problem with redundant variables. Both use the idea of dimensionality reduction. Before a short presentation of these methods, let us assume some ad hoc definitions. The variable is relevant when it affects the response *Y*, individually or in the context with other variables. The variable is irrelevant when it is not relevant. Finally, the variable is redundant when it is relevant, but approximately the same information about the response *Y* is carried by other variables. The formal definitions formulated from the probabilistic point of view are given in the works devoted to feature selection, e.g.: Yu and Liu (2004) and Guyon et al. (2006). Notable is that correlation between predictors does not necessary mean redundancy (Guyon et al., 2006: 10).

The first solution of redundant variables problem is to discard them using a variable selection procedure. The feature selection task can be formulated as an optimisation problem. Thus, the idea is to find a subset of variables for which the classifier will return the most accurate classifications. Due to a hard combinatorial problem, a heuristic search is performed to obtain an approximate solution. For a review of various search techniques, see e.g.: Korf (1999). In the filter approach, feature selection works as the pre-processing step. A quality criterion is assumed which is not directly connected with the model. The multivariate filter criteria are formulated in order to maximise the correlation between predictors and the response *Y*, and simultaneously to minimise the correlations between predictors. The example is a group correlation:

$$H(S) = \frac{k \cdot \overline{r}(X_i, Y)}{\sqrt{k + k(k-1) \cdot \overline{r}(X_i, X_j)}}, \tag{2}$$

where: $\overline{r}$ denotes a mean correlation, and *k* is the cardinality of subset *S*. This criterion was used by Hall (2000) in his CFS (*Correlation-based Feature Selection*) algorithm in combination with the *best-first* search strategy. A slightly different approach was proposed by Yu and Liu (2004) in their FCBF (*Fast Correlation-Based Filter*) algorithm. They have developed a two-step procedure where the relevant features are filtered in the first step, and the redundant variables are discarded in the second one. In fact, the algorithm implements the special way of searching for the best feature subset. An analogues method for regression is described in the book (Grabiński, Wydymus, Zeliaś, 1982). Note that in the discrimination task, the correlation measure must be adjusted to the nominal variable *Y*. Usually, entropy based measures are used. Hall (2000) as well as Yu and Liu (2004) use symmetrical uncertainty:

$$SU(Y, X) = 2 \cdot \frac{H(Y) + H(X) - H(Y, X)}{H(Y) + H(X)}, \tag{3}$$

where H is the entropy measure:

$$H(X) = -\sum_i P(X = x_i) \cdot \log_2 P(X = x_i) \,. \tag{4}$$

As these measures assume both nominal variables, the predictors are previously discretised.

Instead of removing redundant variables, one can aggregate the information carried by these variables by constructing synthetic variables from them. This approach consists of two steps. First, a clustering algorithm divides the variables into groups. Then, a synthetic variable is constructed in each group. Various formulas of aggregating the features can be proposed. We consider the simplest way, namely the linear combination:

$$A = w_1 X_{i_1} + ... + w_k X_{i_k} \,, \tag{5}$$

where k is the number of variables in the cluster. The grouped variables can be normalised before aggregation if it is necessary. In our experiment, we have examined two ways of weighting. In the first one, we simply take all weights equal to one. In the other case, we want to assign greater weight to these features which are more correlated with the response Y. Thus, the coefficients are calculated as:

$$w_j = \frac{SU(Y, X_{i_j})}{\sum_{l=1}^{k} SU(Y, X_{i_l})} \,. \tag{6}$$

4. Empirical study

The goal of the experiment is a comparison of two approaches to the problem of redundant variables. We use the artificial dataset from Section 2 and five datasets from the UCI Machine Learning Repository with additional redundant variables which were generated as follows. Consider the ranking of variables obtained by the random forests importance measure. Take first $q = max$ {3, round (10% p)} variables from this ranking (where p is the number of predictors) and construct five correlated variables for each of them according to the same formula as in Section 2. Include these redundant variables in the original dataset.

We have examined two feature selection algorithms and clustering of the features where two formulas of building the synthetic variables have been considered. We have used several packages of R program for the calculations: `cluster`, `clusterSim`, `FSelector`, `Biocomb` and `randomForest`. After reduction of dimensionality, random forests have been run with default settings including: the

number of trees equal to 500 and the number of sampled features equal to square root of p. The classification error rate has been estimated by the average test error using 30 splits. The first considered feature selection algorithm is the CFS. It maximises the criterion given in formula (2), where symmetrical uncertainty performs the role of correlation measure. The search for the optimal feature subset has been carried out according to the *best-first* strategy. The second feature selection algorithm which we have taken under consideration is the FCBF. Due to available function `select.fast.filter` in the package `Biocomb` of R program, the information gain has been chosen as a measure of correlation (i.e. the numerator in formula (3)). Reduction of dimensionality by clustering of the features leaves many possibilities. We have chosen hierarchical clustering implemented in the function `agnes`, which is available in the package `cluster` of R program. This function allows the dissimilarity matrix to be the input. We have calculated the elements of this matrix according to the formula $1-r_{ij}^2$, where r is Spearman's rank correlation coefficient between the i-th and j-th variable. The linking method applied in our experiment is a group average-link where the distance between two clusters is the average of the dissimilarities between the points in one cluster and the points in the other cluster. The final number of clusters has been determined so as to maximise the silhouette index. We have considered two formulas of building the synthetic variables as it has been described in the previous Section 3. We denote by TC1 the case with all weights equal to one, and by TC2 the case of weights determined by symmetrical uncertainty (6).

Table 1. Misclassification rates with standard errors (%) for original datasets and for datasets with redundant variables. The second column contains p-value from the Wilcoxon rank sum test, where random forests test errors are compared

Datasets		**Original**	**Redundant variables included**				
	p-value	**RF**	**RF**	**CFS+RF**	**FCBF +RF**	**TC1+RF**	**TC2+RF**
artificial	0.01074	1.5 (0.2)	2.2 (0.2)	2.3 (0.3)	1.7 (0.2)	1.5 (0.1)	1.3 (0.1)
ionosphere	0.31900	6.5 (0.4)	7.0 (0.3)	7.4 (0.4)	9.0 (0.5)	6.4 (0.4)	6.3 (0.4)
parkinson	0.14180	10.5 (0.7)	11.8 (0.7)	13.0 (0.7)	13.1 (0.9)	10.6 (0.7)	10.2 (0.6)
segmentation	0.03137	2.3 (0.1)	2.7 (0.1)	3.6 (0.1)	2.3 (0.1)	4.0 (0.2)	4.0 (0.2)
sonar	0.03005	18.4 (0.9)	21.3 (0.8)	24.9 (0.8)	27.0 (0.9)	18.4 (0.8)	19.5 (0.8)
wine	0.00006	1.8 (0.2)	4.5 (0.5)	3.6 (0.4)	2.5 (0.3)	2.5 (0.4)	2.4 (0.4)

Source: own calculations

The results are shown in Table 1. As random forests select features inside the learning algorithm, we have included the classification errors of this method which we have run without any pre-processing. Assuming the significance level of 0.05, we observe that adding redundant variables to the datasets has significantly decreased the accuracy of random forests in 4 out of 6 cases. This result proves

that it is reasonable to reduce the dimension of feature space before running random forests. Surprisingly, the popular CFS algorithm has given even worse results than random forests performed without pre-processing. The only exception is the *wine* dataset but other methods have outperformed the CFS due to the fact that the CFS has not eliminated redundant variables. The FCBF algorithm has proved to be the best on the *segmentation* set, leading to the error rate that has been obtained by random forests on the original dataset. On the other hand, in a few cases, the FCBF has given worse results than the CFS. The most promising results have been obtained by clustering of features. In 4 out of 6 cases, this method has recaptured the level of error that random forests achieve on the original datasets without redundant variables. Notable is that all these datasets consist of two classes. Note also that the way of constructing synthetic variables has not influenced significantly the error rate.

5. Conclusions

Random forests that are commonly considered as one of the best "off-the-shelf" classifiers in the world have their drawbacks. We have shown that redundant variables may deteriorate their accuracy. Moreover, the pre-processing step with usually performed filters does not always work. In fact, this approach has failed in most cases in our experiment. We have examined an alternative method to overcome the problem of redundant variables. Instead of eliminating the features, they can be grouped in clusters, and then the groups can be represented by synthetic variables. This approach is a promising tool in dealing with redundant variables. The inconvenience is that this method requires many settings, e.g.: the dissimilarity measure, the linking method in hierarchical clustering, or the way of constructing the synthetic variables. However, having a large enough dataset, one can validate various settings and methods. Based on our results, it seems that the clustering of features is not so efficient in the case of multiclass discrimination. This requires further research.

References

Breiman L. (1996), *Bagging predictors*, "Machine Learning", vol. 24(2), pp. 123–140.

Breiman L. (2001), *Random forests*, "Machine Learning", vol. 45, pp. 5–32.

Freund Y., Schapire R.E. (1996), *Experiments with a new boosting algorithm*, Proceedings of the 13th International Conference on Machine Learning, Morgan Kaufmann, San Francisco.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Grabiński T., Wydymus S., Zeliaś A. (1982), *Metody doboru zmiennych w modelach ekonometrycznych*, Państwowe Wydawnictwo Naukowe PWN, Warszawa.

Granitto P. M., Furlanello C., Biasioli F., Gasperi F. (2006), *Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products*, "Chemometrics and Intelligent Laboratory Systems", vol. 83(2), pp. 83–90.

Gregorutti B., Michel B., Saint-Pierre P. (2017), *Correlation and variable importance in random forests*, "Statistics and Computing", vol. 27, issue 3, pp. 659–678.

Guyon I., Gunn S., Nikravesh M., Zadeh L. (2006), *Feature Extraction: Foundations and Applications*, Springer, New York.

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.

Hapfelmeier A., Ulm K. (2013), *A new variable selection approach using Random Forests*, "Computational Statistics and Data Analysis", vol. 60, pp. 50–69.

Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning: Data Mining. Inference and Prediction*, 2nd edition, Springer, New York.

Korf R.E. (1999), *Artificial intelligence search algorithms*, [in:] M.J. Atallah, *Algorithms and Theory of Computation Handbook*, CRC Press, Boca Raton–London–New York–Washington.

Kursa M.B., Rudnicki W.R. (2010), *Feature selection with the Boruta package*, "Journal of Statistical Software", vol. 36, issue 11, pp. 1–13, http://www.jstatsoft.org/v36/i11/ [accessed: 15.02.2018].

Toloşi L., Lengauer T. (2011), *Classification with correlated features: unreliability of feature ranking and solutions*, "Bioinformatics", vol. 27, issue 14, pp. 1986–1994, https://doi.org/10.1093/bioinformatics/btr300.

Ye Y., Wu Q., Zhexue Huang J., Ng M.K., Li X. (2013), *Stratified sampling for feature subspace selection in random forests for high dimensional data*, "Pattern Recognition", vol. 46(3), pp. 769–787, https://doi.org/10.1016/j.patcog.2012.09.005.

Yu L., Liu H. (2004), *Efficient feature selection via analysis of relevance and redundancy*, "Journal of Machine Learning Research", no. 5, pp. 1205–1224.

Problem zmiennych redundantnych w metodzie lasów losowych

Streszczenie: Lasy losowe są obecnie jedną z najchętniej stosowanych przez praktyków metod klasyfikacji wzorcowej. Na jej popularność wpływ ma możliwość jej stosowania bez czasochłonnego, wstępnego przygotowywania danych do analizy. Las losowy można stosować dla różnego typu zmiennych, niezależnie od ich rozkładów. Metoda ta jest odporna na obserwacje nietypowe oraz ma wbudowany mechanizm doboru zmiennych. Można jednak zauważyć spadek dokładności klasyfikacji w przypadku występowania zmiennych redundantnych. W artykule omawiane są dwa podejścia do problemu zmiennych redundantnych. Rozważane są dwa sposoby przeszukiwania w podejściu polegającym na doborze zmiennych oraz dwa sposoby konstruowania zmiennych syntetycznych w podejściu wykorzystującym grupowanie zmiennych. W eksperymencie generowane są liniowo zależne predyktory i włączane do zbiorów danych rzeczywistych. Metody redukcji wymiarowości zwykle poprawiają dokładność lasów losowych, ale żadna z nich nie wykazuje wyraźnej przewagi.

Słowa kluczowe: lasy losowe, zmienne redundantne, dobór zmiennych, taksonomia cech

JEL: C1, C38, C52

Received: 2018-02-18; verified: 2018-07-27. Accepted: 2018-09-24