THE INFLUENCE OF UNBALANCED ECONOMIC DATA ON FEATURE SELECTION AND QUALITY OF CLASSIFIERS

Mariusz Kubus, Ph.D.

Opole University of Technology

Faculty of Production Engineering and Logistic

Department of Mathematics and Applied Computer Science

Sosnkowskiego 31, 45-272 Opole, Poland

e-mail: m.kubus@po.edu.pl

ORCID: 0000-0002-6602-2742

Received 19 November 2019, Accepted 31 March 2020

Abstract

Research background: The successful learning of classifiers depends on the quality of data. Modeling is especially difficult when the data are unbalanced or contain many irrelevant variables. This is the case in many applications. The classification of rare events is the overarching goal, e.g. in bankruptcy prediction, churn analysis or fraud detection. The problem of irrelevant variables accompanies situations where the specification of the model is not known *a priori*, thus in typical conditions for data mining analysts.

Purpose: The purpose of this paper is to compare the combinations of the most popular strategies of handling unbalanced data with feature selection methods that represent filters, wrappers and embedded methods.

Research methodology: In the empirical study, we use real datasets with additionally introduced irrelevant variables. In this way, we are able to recognize which method correctly eliminates irrelevant variables.

Results: Having carried out the experiment we conclude that over-sampling does not work in connection with feature selection. Some recommendations of the most promising methods also are given.

Novelty: There are many solutions proposed in the literature concerning unbalanced data as well as feature selection. The innovative field of our interests is to examine their interactions.

Keywords: classifiers, class unbalance, sensitivity, feature selection, resampling

JEL classification: C1, C38, C52

Introduction

The problem of unbalanced datasets in supervised classification occurs when examples of one class are rare compared to other classes. This happens quite often even in the case of binary classification, which is considered in this paper. Just to mention applications such as bankruptcy prediction, churn analysis, the prediction of positive responses in direct marketing campaigns, or fraud detection. When examples of one class are rare, the classifiers tend to accurate the prediction of the opposite class. This means that most of the rare examples are misclassified while the majority class is recognized quite well. In extreme cases, the models can classify all objects into the majority class. The explanation of this phenomenon depends on the learning method. In logistic regression, the posterior probabilities of minority class are underestimated (King, Zeng, 2001). In turn, recursive partitioning learning algorithm minimizes overall the error regardless of class. Note that the correct classification of rare examples is of particular importance in the mentioned applications. The model that accurately predicts the bankruptcy of enterprises contributes to investment risk reduction. Detection of frauds, which are a marginal part of overall transactions, protects against losses. Moreover, it affects customer trust and the bank's image. In a churn analysis, resigning customers are usually a small fraction of all customers. A lot of attention is paid to them, because customer retention is much cheaper than getting a new one. That is why particular offers are prepared to unsure clients and they are often individually suited to them. The cost of direct marketing action decreases when more potentially resigning customers are identified.

The unbalanced learning problem has gained a great deal of interest recently. Providing a complete review of the literature falls beyond the scope of this paper. The overview of existing solutions with wide references can be found in the works of (Chawla, Japkowicz, Kołcz, 2004; Weiss, 2004; Galar, Fernandez, Barrenechea, Bustince, Herrera, 2011; Longadge, Dongre, Malik, 2013; Haixiang et al., 2017). These solutions can be divided into two groups, where changes are introduced at a level of data or learning algorithm. Note that this classification is not separable. Consider the balancing methods of resampling. They are usually used in the pre-processing step to prepare a new training set, which would be balanced. However, the same techniques can be applied inside a learning algorithm, namely for each base model in ensemble learning. Balanced random forests can be cited as an example (Chen, Liaw, Breiman, 2004), where the training sample is balanced in each iteration.

It should be noted that the assessment of models, obtained from unbalanced data, requires other quality metrics than the overall classification error. The most popular are: sensitivity (a.k.a.

recall), specificity, precision, AUC measure (which is a mean of accuracy calculated for both classes separately), or *F*-measure that combines sensitivity and precision (Fawcett, 2006). All of them distinguish incorrect classifications of both, minority and majority class. The extensive study of model assessment in the context of unbalanced data is given in (Japkowicz, Shah, 2011).

The next problem of building the models with a high accuracy of classification are irrelevant variables, i.e. those that have no impact on the response variable (Guyon, Gunn, Nikravesh, Zadeh, 2006). The relevancy of a variable is used to be defined as individual or contextual with other variables. Irrelevant variables can lead to over fitting, decreasing the generalization ability of the models. In addition, the accuracy of the model parameter's estimation decreases with an increase of dimension (*curse of dimensionality*). Frequently, the goal of analysis is the discovery of unknown relations in big datasets, when a researcher does not have sufficient prior knowledge about which features really influence a dependent variable. Therefore, a feature selection is an important stage that may be formulated as an optimisation problem, namely the task of finding a subset of predictors that would give the best classifier (Tsamardinos, Aliferis, 2003).

The modeling of classifiers is particularly difficult when the described problems occur simultaneously. We suppose that unbalanced data influence the feature selection process. The goal of this paper is to examine the relation between balancing techniques and feature selection in order to build finally an optimal classifier. We consider seven techniques of handling unbalanced data, and seven feature selection methods, which represent three main approaches to this problem: filters, wrappers and embedded methods. In the presented empirical study we use real datasets with artificially added irrelevant variables. This made it possible to observe the relationship between the balancing of training sets and variable selection. Our study allowed us to draw a few conclusions and recommendations on the use of discussed methods.

The rest of this paper is organized as follows. Section 1 shortly presents the most popular strategies of handling unbalanced data. Section 2 is devoted to feature selection problem. The setup and the results of our experiment are presented in Section 3. Then in Section 4 we include the concluding remarks.

1. How to handle unbalanced data?

Before presenting the popular ways of dealing with unbalanced data, we consider some simple model quality measures that distinguish between types of misclassifications. We introduce terminology and notations commonly used in the literature (Fawcett, 2006). The examples from

minority class and majority class are called positive and negative respectively. The results of the classification obtained by a given model are summarised in the confusion matrix, where the numbers of correct classifications are denoted by TN (*true negative*) and TP (*true positive*) for majority and minority class respectively, and the numbers of incorrect classifications are denoted by FP (*false positive*) for misclassified negative examples or FN (*false negative*) for misclassified positive ones. A simple measure which focuses on rare events is the accuracy calculated for minority class only $TPR = TP / (TP + FN)$. It is called a true positive rate or sensitivity. In turn, the accuracy of the classification of negative examples $TNR = TN / (TN + FP)$ is called true negative rate, or specificity. The balance between these two kinds of accuracies is reflected in the AUC measure, which is simply their arithmetic mean.

Two general approaches to solving the problem of unbalance are changes in data structure or the modifications of learning algorithms. The first is to prepare a new, balanced training set. The simplest balancing techniques use random resampling. Depending on the fraction of the minority class and of data size, one can use under-sampling, over-sampling or a mixed approach. Under-sampling decreases the size of the training set leaving all examples from the minority class and picking only a small fraction of examples from the majority class. This can be beneficial in the case of processing large datasets, but on the other hand, important information can be removed from the data, affecting the quality of the model. In turn, over-sampling leaves all examples from the original dataset and additionally it replicates randomly examples from the minority class. A mixed approach follows both scenarios. More sophisticated methods of data balancing use information content from a training set, generally in the form of conditional distributions or nearest neighbors. G. Menardi and N. Torelli (2014) proposed the ROSE algorithm, which generates new synthetic examples so that the classes were equally represented. For this purpose they use the density function $f(\mathbf{x}|Y = y_i)$, which is approximated by the kernel estimate:

$$\hat{f}(\mathbf{x}|Y = y_i) = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{H_j}(\|\mathbf{x}, \mathbf{x}_i\|) \quad (1)$$

where j indicates a class, n_j is its size and H_j is a matrix of scale parameters in the chosen class. Under- as well as over-sampling can be performed to reach a pre-defined number of observations in both classes. The algorithm is partially random because the objects $(\mathbf{x}_i, y_i)$ of the original dataset are drawn randomly with the same class probabilities. The most popular algorithm that uses distances between objects is SMOTE (Chawla et al., 2002). SMOTE creates synthetic examples in the neighbourhood of minority class points, up to reaching approximately

an equal cardinality of classes. The k nearest neighbors of the minority class are found for each positive example $\mathbf{x}$. One of them $NN^*(\mathbf{x})$ is drawn randomly and the new object is created by shifting the point $\mathbf{x}$ towards the nearest neighbor $NN^*(\mathbf{x})$. The distance of that shifting is set randomly. The number k as well as a number of drawn neighbors $NN^*(\mathbf{x})$ are the parameters of the algorithm.

The second approach of dealing with unbalanced data is a modification of the learning algorithm. In practice, the change of estimation criterion is not an easy solution. It requires thorough knowledge on the classification method and programming skills. Assigning different costs to the incorrect classification also is usually problematic due to insufficient knowledge on the examined phenomenon. The simplest way is to run learning an algorithm for original data and to shift the classification cut-off point for posterior probabilities.

The techniques described above may be used in combination with certain methods of cleaning the data, which focus on the noisy examples or overlapped regions. As an example we can present I. Tomek links method (Tomek, 1976). This method searches for pairs of points that are the nearest neighbours to each other but represent different classes. Then both or one example from the majority class is removed. The result is that all pairs of nearest neighbours in the data belong to the same class.

2. Feature selection

The task of the feature selection can be formulated as a problem of combinatorial optimisation. Given some quality criterion $Q = Q(S)$ we are to search for the feature subset $S \subseteq \mathbf{X}$, so that the function $Q(S)$ reaches its optimum. Due to the relationship between the search process and the estimation of model parameters, feature selection methods are currently grouped into three approaches: filters, wrappers and embedded methods (Guyon et al., 2006).

The filters work as a pre-processing step and the criterion Q is not directly connected with model quality. The most popular is an evaluation of the individual impact of a predictor on a dependent variable. This approach is fast, easy to implement and the process of the search is reduced to ranking the variables. This may be performed by statistical significance tests. In the case of a qualitative response variable, the Wilcoxon rank sum test or chi-squared test can be applied, depending on the type of predictor. The significance level decides which variables are treated as relevant. The next popular group of filter criteria use entropy measure:

$$H(X) = -\sum_i P(x_i) \cdot \log_2 P(x_i) \quad (2)$$

which express the information content in the distribution. The measure (2) assumes qualitative or discrete variables. The continuous variables should be discretised. One of the most popular entropy based measure is a normalized version of mutual information, which is called symmetrical uncertainty:

$$SU(Y, X) = 2 \cdot \left(1 - \frac{H(Y, X)}{H(Y) + H(X)} \right) \quad (3)$$

Note, that applying this criterion results only in the ranking of variables, and it is necessary to set a certain relevance threshold. A relatively fast algorithm that eliminates irrelevant as well as redundant variables was introduced by L. Yu and H. Liu (2004). In the first stage of their FCBF algorithm, the relevant variables are filtered using a correlation measure, which is adjusted to the nominal dependent variable Y . Originally it was symmetrical uncertainty. In the second stage, the redundant variables are removed. The variable is considered as redundant if it is correlated with the other predictor, which has a greater correlation with Y . To state if there is a correlation between features or not, a threshold value is introduced. A separate group of filters represents a multidimensional approach. Some of them use distances between objects. The classical example is the ReliefF algorithm (Kononenko, 1994). It assigns weights to the variables in the iterative procedure, in which example $\mathbf{x}$ is sampled from a training set and its k nearest neighbours from the same class (*nearest hits NH*) are searched, as well as k nearest neighbours from the opposite class (*nearest misses NM*). Then the weight of the variable is updated according to the coordinate differences between $\mathbf{x}$ and its nearest neighbours:

$$w_j \leftarrow w_j - \frac{1}{k} \sum_{i=1}^k |x_j - NH_i(\mathbf{x})_j| + \frac{1}{k} \sum_{i=1}^k |x_j - NM_i(\mathbf{x})_j| \quad (4)$$

The algorithm has two input parameters: k and the number of iterations. The recommendations of their settings are given e.g. in (Kubus, 2016). The obtained weights give a feature ranking, thus the cut-off point for relevance has to be set yet.

The second group of feature selection methods is constituted by so called wrappers. Model assessment is employed for the evaluation of feature subsets. Searching the space of all subsets is performed as an outer loop of the learning algorithm. This approach can have a high computational burden, even for heuristic searches. The most popular representative of wrappers is the stepwise procedure, which may be applied in two directions. Forward selection adds a variable to the current subset in each iteration so that it optimises the model quality criterion. On the contrary, backward elimination discards one variable from the current subset in each

iteration. This second variant starts running with a full set of predictors and is much more computationally expensive, especially when the complexity of learning algorithms is high. To accelerate this searching process recursive feature elimination was proposed in the context of support vectors machines (Guyon, Weston, Barnhill, Vapnik, 2002). M. Kubus (2015) emphasizes that recursive feature elimination is a general procedure, which can work with many discrimination or regression methods. In the context of logistic regression, this algorithm can be found to be popular and is willingly used by practitioners with the technique of iterative eliminating variables corresponding with insignificant coefficients.

The last approach to the task of feature selection is to place this mechanism inside the learning algorithm (*embedded methods*). Searching for the best subset and learning a model is performed simultaneously. The examples of such methods are classification trees (also in ensemble learning) and regularised logistic regression. In this second method, the penalty component $P(\mathbf{b})$ is introduced to the estimation criterion:

$$\hat{\mathbf{b}} = \arg \min_{\mathbf{b}} (-2 \ln L(\mathbf{b}) + \lambda \cdot P(\mathbf{b})) \quad (5)$$

where $L(\mathbf{b})$ is a likelihood function. The higher the coefficient values, the greater the penalty value. This results in the shrinking of coefficients to zero. In extreme cases they can be equal to zero, which means the feature selection. The penalty term called elastic net:

$$P_{\alpha}(\mathbf{b}) = \sum_{j=1}^p \left((1 - \alpha) b_j^2 + \alpha |b_j| \right) \quad (6)$$

was proposed by H. Zou and T. Hastie (2005) and it combines historically earlier propositions: the ridge regression and lasso. The penalty parameter λ influence the amount of shrinking. In practice, several models are estimated for different values of λ and finally this one is chosen which minimizes the classification error or information criterion.

3. Empirical study

The unbalanced data we have used in the empirical study come from the UCI Machine Learning Repository (Dua, Graff, 2019). Our intention was a choice of economic data that represent popular applications, such as the prediction of potential customers in a direct marketing campaign, churn analysis, and bankruptcy prediction. A summary is shown in Table 1. Each of them represents a binary classification problem. The nominal variables in the *Bank marketing* dataset were transformed into binary variables for the purpose of logistic

regression models. Note that there were a lot of missing data in the set *Polish bankruptcy*. We removed such rows so that this additional problem would not influence the results. There are still 4,769 examples left in this dataset. Because the goal of our study was to examine the relationship between an unbalance problem and feature selection, we added irrelevant variables to the original data. The number of irrelevant variables included in each set was approximately 25% of the number of the original variables. They were independently generated from the normal distribution with a zero correlation. The exception was *Bank marketing* dataset where half of the irrelevant variables were binary with equal fractions of zeros and ones. This was because this dataset consisted of many dummy variables.

Table 1. Datasets summary

Datasets	# Examples	# Predictors (with dummy)	Fraction of minority class (%)
Bank-marketing	4,521	16 (42)	11.5
Churn	5,000	18	14.1
Polish bankruptcy 4	4,769	64	2.5

Source: UCI Machine Learning Repository.

Due to the importance of economic interpretation we considered the logistic regression model, which is also characterized by a rapid estimation process and it is sufficient when a class structure is not very complicated. On the other hand, we compared the results with random forests. These classifiers have very high prediction accuracy and perform automatic feature selection, but they act like a black box, which is a disadvantage when interpretation is of particular interest. An additional advantage of random forests is the low number of hyperparameters. When building trees without pruning, it is enough to determine their number and the number of variables randomly drawn at the tree nodes. Setting these parameters is suggested in a source article (Breiman, 2001). In our investigation, the number of variables randomly selected in the nodes is approximately a square root of the number of predictors. The number of trees in the forest is set as 200. The model quality measures are estimated via a 10-fold cross-validation. It is noteworthy that the split into 10 folds were made only once so that all methods could be compared on the same training and test sets.

In our research, we combined 7 techniques of dealing with the unbalance problem with 7 techniques of feature selection. Below we present the settings and notations that are used in the tables and figures.

- US:** random under-sampling, where all positive examples remain in the new training set and an equinumerous set of negative examples is sampled without replacement.
- OS:** random over-sampling, where all examples remain in the new training set and additional set of positive examples is sampled with a replacement so that the classes were equinumerous.
- MIX:** the combination of under- and over-sampling. The half number of negative examples is sampled without replacement and then a positive class is over-sampled to obtain equinumerosity.
- ROSE:** balancing the data as in the mixed approach but instead of random sampling the new synthetic observations are generated from the conditional density, which is approximated by a normal product function centered at each of the $\mathbf{x}_i$ with the diagonal covariance matrix H_k , where H_k is the asymptotically optimal smoothing matrix under the assumption of multivariate normality.
- SMOTE:** the number of nearest neighbours during the sampling process was set as 5. The positive class was over-sampled to obtain equinumerosity.
- Cut:** a direct learning on original unbalanced training sets and the use of shifted cut-off point for classification. The cut-off point was set as the fraction of the minority class in a training set.
- TL-Cut:** as above but the data was pre-processed with the Tomek links technique.
- TEST:** filtering of the features using statistical tests (Wilcoxon rank sum test for quantitative features and chi-squared test for qualitative). The significance level was set as 0.1.
- SU:** filtering of the features with the symmetrical uncertainty measure. We set the threshold of feature relevance as 0.01. The quantitative variables were discretised following MDL method proposed by U. Fayyad and K. Irani (1993).
- ReliefF:** filtering of the features according to their weights assigned by the ReliefF algorithm. The number of iterations was set as a half the cardinality of the training set, and the number of nearest neighbours was approximately $N_p^{0.25}$ (where N_p denotes the number of positive examples in the training set). The threshold of feature relevance was set as 0.05.
- FCBF:** the correlation measure in the FCBF algorithm is symmetrical uncertainty. The threshold of feature relevance as well as of inter-correlation between predictors was set as 0.01. Discretisation as in SU.

RFE: recursive feature elimination where all variables corresponding to insignificant coefficients at a level of 0.05 are eliminated in each iteration.

STEP: forward stepwise regression. As the optimisation criterion we used the Bayesian information criterion BIC.

RLR: regularized logistic regression with elastic net penalty. Alpha parameter in eq. (6) was set as 0.9 so that to assign more weight to the term with absolute value, which decides about feature selection effect. Penalty parameter λ was determined according to the minimal value of the Bayesian information criterion BIC.

The first stage of our research was to examine the impact of balancing techniques on feature selection process. We also considered feature selection on unbalanced training sets for the purpose of classification according to the shifted cut-off point. The numbers of undetected irrelevant variables are presented in Figure 1. For each technique of handling unbalance data, these numbers were summed over 7 feature selection methods, 3 datasets and 10 folds of cross-validation, thus for 210 models. The largest numbers of selected irrelevant variables were obtained in the case of over-sampling techniques, regardless of whether they were random or “intelligent”, i.e. using information contained in the data. The worst results that can be observed are for the SMOTE algorithm and for random over-sampling. Note, that in both cases the minority class is replicated up to the cardinality of the majority class. The large number of irrelevant variables not recognized by feature selection techniques is also the same for the mixed form of over- and under-sampling, even when using information about the distributions,

Figure 1. Numbers of irrelevant variables introduced to the models over a 10-fold cross-validation and 7 feature selection techniques for three datasets: *bank marketing* (black), *churn* (dark grey) and *Polish bankruptcy* (bright grey)

Source: own calculations.

as in the ROSE algorithm. The best result we observed was for under-sampling, despite critical comments regarding the loss of information in this method. We note that the worst results were obtained for the *Polish bankruptcy* dataset. This may be connected with the lowest fraction of positive examples in this dataset. Summarizing the above comments one can conclude that the only a sensible balancing technique from the feature selection point of view is under-sampling. The feature selection does not work in the case of duplicating information content in the data or adding artificial examples, even if these techniques are based on a strong theoretical basis.

The feature selection in unbalanced training sets (Cut & TL-cut) deserves a separate discussion. In these cases, the numbers of irrelevant variables selected are comparable to under-sampling, except for the *Polish bankruptcy* dataset. Although slightly worse results were obtained after cleaning the data with Tomek links, but this mainly concerns to the *Bank marketing* dataset. We can suppose that this is because of many binary variables in this dataset. For these reasons, we will incorporate these techniques in further research. Table 2 shows the numbers of undetected irrelevant variables that correspond to the methods of feature selection.

Table 2. The medians of irrelevant variables undetected in the cross-validation

Unbalance solution	TEST	SU	ReliefF	FCBF	RFE	STEP	RLR
Bank marketing							
US	0	0	0	0	0	0	0.5
Cut	0	0	2	0	0	0	0
TL-cut	0	0	4	0	0	0	0
Churn							
US	1	0	0	0	2	0	2
Cut	1	0	0	0	1	0	0
TL-cut	0	0	0	0	1	0	1
Polish bankruptcy							
US	2	0	1	0	0	0	0
Cut	1	0	16	0	0	0	0
TL-cut	2	0	16	0	0	0	0

Source: own calculations.

The next stage of our research was to compare the model quality measures for the combinations of feature selection and strategies of handling unbalance data. The results are presented in Table 3. To assess the significance of the differences between the means of AUC (or TPR) we conducted the Friedman rank sum test, separately for each technique of handling

unbalance and for each dataset. The null hypothesis was rejected in all cases, thus a post hoc analysis was carried out with the use of the Nemenyi test. Due to the large number of results, we have presented them in the summary version. We made the scoring of feature selection methods according to the following principles. The method is given a point if its mean of AUC (or TPR) does not differ significantly from the best result. The level of significance has been set as 0.1. The results summed over three datasets are presented in (Table 4). The winners in this scoring are SU, STEP and RLR. The first two methods are simultaneously among the top three from the point of view of identifying irrelevant variables (see Table 2). The third method, FCBF has failed

Table 3. Model quality metrics estimated via a 10-fold cross-validation
(standard errors in brackets)

	US		Cut		TL-cut	
	AUC	TPR	AUC	TPR	AUC	TPR
Bank marketing						
TEST	0.804 (0.007)	0.778 (0.014)	0.804 (0.009)	0.773 (0.017)	0.809 (0.007)	0.808 (0.013)
SU	0.778 (0.008)	0.721 (0.013)	0.792 (0.007)	0.744 (0.014)	0.797 (0.006)	0.778 (0.012)
ReliefF	0.723 (0.005)	0.633 (0.013)	0.769 (0.004)	0.758 (0.009)	0.766 (0.006)	0.775 (0.015)
FCBF	0.768 (0.006)	0.700 (0.011)	0.758 (0.004)	0.671 (0.007)	0.765 (0.005)	0.765 (0.005)
RFE	0.803 (0.007)	0.773 (0.014)	0.810 (0.007)	0.800 (0.013)	0.817 (0.007)	0.825 (0.013)
STEP	0.804 (0.007)	0.778 (0.013)	0.797 (0.009)	0.758 (0.016)	0.806 (0.007)	0.795 (0.014)
RLR	0.800 (0.008)	0.754 (0.020)	0.800 (0.007)	0.766 (0.016)	0.805 (0.007)	0.805 (0.016)
Churn						
TEST	0.766 (0.009)	0.766 (0.023)	0.764 (0.008)	0.768 (0.021)	0.769 (0.007)	0.803 (0.016)
SU	0.766 (0.008)	0.765 (0.019)	0.760 (0.011)	0.764 (0.023)	0.755 (0.010)	0.795 (0.022)
ReliefF	0.741 (0.009)	0.730 (0.022)	0.766 (0.008)	0.773 (0.018)	0.767 (0.009)	0.808 (0.020)
FCBF	0.748 (0.008)	0.738 (0.020)	0.744 (0.009)	0.736 (0.023)	0.745 (0.008)	0.769 (0.022)
RFE	0.670 (0.013)	0.653 (0.029)	0.677 (0.014)	0.619 (0.028)	0.672 (0.013)	0.658 (0.024)
STEP	0.776 (0.009)	0.787 (0.023)	0.763 (0.010)	0.770 (0.021)	0.762 (0.009)	0.806 (0.020)
RLR	0.761 (0.011)	0.766 (0.019)	0.767 (0.009)	0.789 (0.022)	0.768 (0.009)	0.818 (0.021)
Polish bankruptcy						
TEST	0.703 (0.044)	0.713 (0.050)	0.751 (0.034)	0.693 (0.087)	0.706 (0.037)	0.575 (0.100)
SU	0.774 (0.022)	0.767 (0.043)	0.763 (0.026)	0.737 (0.072)	0.753 (0.030)	0.711 (0.079)
ReliefF	0.693 (0.030)	0.699 (0.059)	0.650 (0.022)	0.655 (0.047)	0.659 (0.022)	0.679 (0.048)
FCBF	0.721 (0.019)	0.736 (0.039)	0.698 (0.017)	0.849 (0.034)	0.693 (0.017)	0.872 (0.033)
RFE	0.568 (0.030)	0.251 (0.104)	0.722 (0.016)	0.817 (0.029)	0.739 (0.020)	0.870 (0.036)
STEP	0.768 (0.021)	0.810 (0.044)	0.770 (0.023)	0.805 (0.041)	0.767 (0.023)	0.799 (0.043)
RLR	0.706 (0.033)	0.661 (0.092)	0.719 (0.034)	0.800 (0.100)	0.697 (0.039)	0.736 (0.128)

Source: own calculations.

in the comparisons of AUC and TPR. This is because it discards features that are too radical, removing in this way important information. Regularised logistic regression is among the winners, although it has not eliminated irrelevant variables perfectly. We guess that this could be because of the low values of coefficients that correspond with irrelevant variables. Noteworthy is that TEST failed only once, namely for TL-cut in the *Polish bankruptcy* dataset. This method is simple and fast, thus it deserves for a closer look. Experiments with a more radical level of confidence would be welcomed in future work. The RFE algorithm has yielded great results for the *Bank marketing* dataset but it has failed for the other datasets.

Table 4. Scoring of feature selection methods due to the significance of differences between the mean values of AUC and TPR

	US		Cut		TL-cut		Mean	
	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
TEST	3	3	3	3	3	2	3.00	2.67
SU	3	3	3	3	3	3	3.00	3.00
ReliefF	1	2	1	2	1	1	1.00	1.67
FCBF	1	1	2	2	2	2	1.67	1.67
RFE	1	1	2	2	2	2	1.67	1.67
STEP	3	3	3	3	3	3	3.00	3.00
RLR	3	3	3	3	3	3	3.00	3.00

Source: own calculations.

In the next stage of the empirical study we compared the results obtained with random forests (RF). We ran this algorithm on original, unbalanced data as well as on those cleaned with Tomek links (TL RF). In addition, balanced random forests (BRF) were also compared. This method is the adaptation of random under-sampling to the ensemble learning. It means that resampling is performed separately for each base model (Chen et al., 2004). The results are presented in Table 5. In the *Bank marketing* dataset, there are no significant differences for AUC but TPR for BRF are significantly lower than for other methods. In the *Churn* dataset, AUC for BRF is significantly higher, but at the same time TPR for BRF is significantly lower. For this reason, classic random forests would be preferred to their balanced version. Obviously, there is a shifted threshold for classification in the classic version. In the *Polish bankruptcy* dataset, the differences between the means are significant neither for AUC nor for TPR. As above for logistic regression, we used the Friedman test and pairwise comparisons using the Nemenyi test. The significance level was set as 0.05. Note that the obtained results are better than in

logistic regression. Thus, random forest can be recommended if comprehensibility is not of particular importance.

Table 5. Randoforests assessment. The quality metrics estimated via a 10-fold cross-validation (standard errors in brackets)

Method	Bank marketing		Churn		Polish bankruptcy	
	AUC	TPR	AUC	TPR	AUC	TPR
BRF	0.829 (0.007)	0.865 (0.014)	0.895 (0.006)	0.858 (0.011)	0.808 (0.017)	0.872 (0.031)
RF	0.836 (0.007)	0.896 (0.015)	0.878 (0.005)	0.870 (0.011)	0.774 (0.018)	0.854 (0.036)
TL RF	0.833 (0.008)	0.913 (0.016)	0.869 (0.005)	0.871 (0.011)	0.777 (0.021)	0.866 (0.044)

Source: own calculations.

Conclusions

Unbalanced data as well as irrelevant variables constitute a serious problem in modelling useful classifiers. Both problems have been taken into account in our empirical study, and the results that have been obtained allow us to draw some conclusions. Over-sampling does not work in the context of feature selection. It is meaningful that even “intelligent” versions, that use information from the data, have failed in the presence of irrelevant variables. Besides, we can note that the lower the fraction of the minority class, the more difficult the task of the feature selection is. While in the churn analysis, the fraction of resigning customers is often at the level of a dozen or so per cent, then in forecasting bankruptcy it is not the whole 5 per cent. Depending on the prediction horizon (one or two years) it is from 0.6 to 4.6% in the Polish economy (Pociecha, Pawełek, Baryła, Augustyn, 2014). Note, that the fraction of the minority class in credit card fraud detection is hardly 0.1%.

The superiority of random under-sampling performance is actually beneficial in processing large datasets. The learning of classifiers with the use of original (or cleaned by Tomek links) datasets may also be recommended, and these approaches at least do not increase the sizes of training sets. Among the most promising feature selection methods for unbalanced data we can list: filtering by symmetrical uncertainty measure, forward stepwise regression and regularised logistic regression. The first method is model free, fast and can be especially recommended for large datasets. Note that this filter could be applied also in multiclass classification problems. In turn, the forward stepwise procedure, which have occurred to be perfect in the recognition of irrelevant features, may be computationally expensive in high dimensional problems. Again, combining the stepwise procedure with under-sampling can be

advantageous. It seems that due to the computational cost, the subset of features filtered by symmetrical uncertainty can be set as the starting point in the stepwise search. We consider this to be one of the directions of future work. Using tests of statistical significance is also worthy of further research. In our study this has yielded nearly optimal results. Nevertheless, if interpretation is not very important, using random forests is definitely recommended. This method performs automatic feature selection and does not require the transformation of qualitative features into dummy variables in the pre-processing step. Let us also note that the results obtained after applying the Tomek links were usually comparable with those obtained for original data. As this cleaning technique is not very time consuming it is worthy of being taken under consideration during analysis.

References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chawla, N.V., Japkowicz, N., Kołcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6 (1), 1–6.
- Chen, C., Liaw, A., Breiman, L. (2004) Using random forest to learn imbalanced data. *University of California, Berkeley*, 110, 1–12.
- Dua, D., Graff, C. (2019). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from: <http://archive.ics.uci.edu/ml> (17.06.2019).
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.
- Fayyad, U., Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp. 1022–1027).
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42 (4), 463–484.
- Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (2006). *Feature Extraction: Foundations and Applications*. New York: Springer.

- Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46, 389–422.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239.
- Japkowicz, N., Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.
- King, G., Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.
- Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF, *Proceedings of European Conference on Machine Learning* (pp. 171–182).
- Kubus, M. (2015). Rekurencyjna eliminacja cech w metodach dyskryminacji. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 384, *Taksonomia*, 24, 154–162. DOI: 10.15611/pn.2015.384.16.
- Kubus, M. (2016). Lokalna ocena mocy dyskryminacyjnej zmiennych. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 427, *Taksonomia* 27, 143–152. DOI: 10.15611/pn.2016.427.15.
- Longadge, R., Dongre, S.S., Malik, L. (2013). Class Imbalance Problem in Data Mining: Review. *International Journal of Computer Science and Network*, 2 (1), 83–87.
- Menardi, G., Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28, 92–122.
- Pociecha, J., Pawełek, B., Baryła, M., Augustyn, S. (2014). *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*. Kraków: Fundacja Uniwersytetu Ekonomicznego w Krakowie.
- Tomek, I. (1976). Two modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics*, 6, 769–772.
- Tsamardinos, I., Aliferis, C.F. (2003). Towards principled feature selection: relevancy, filters and wrappers. In *Proceedings of the Workshop on Artificial Intelligence and Statistics*.
- Weiss, G. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations*, 6 (1), 7–19.
- Yu, L., Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5, 1205–1224.
- Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67 (2), 301–320.