sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2020, Vol. 24, No. 1
ISSN 1507-3866; e-ISSN 2449-9994

EVALUATION OF RESAMPLING METHODS IN THE CLASS UNBALANCE PROBLEM

Mariusz Kubus
Opole University of Technology, Opole, Poland
e-mail: m.kubus@po.edu.pl
ORCID: 0000-0002-6602-2742

DOI: 10.15611/eada.2020.1.04
JEL Classification: C1, C38, C52

Abstract: The purpose of many real world applications is the prediction of rare events, and the training sets are then highly unbalanced. In this case, the classifiers are biased towards the correct prediction of the majority class and they misclassify a minority class, whereas rare events are of the greater interest. To handle this problem, numerous techniques were proposed that balance the data or modify the learning algorithms. The goal of this paper is a comparison of simple random balancing methods with more sophisticated resampling methods that appeared in the literature and are available in R program. Additionally, the authors ask whether learning on the original dataset and using a shifted threshold for classification is not more competitive. The authors provide a survey from the perspective of regularized logistic regression and random forests. The results show that combining random under-sampling with random forests has an advantage over other techniques while logistic regression can be competitive in the case of highly unbalanced data.

Keywords: class unbalance, resampling, regularized logistic regression, random forests.

1. Introduction

In many applications of supervised classification the datasets are unbalanced. This means that the number of objects from one class is greatly outnumbered by other classes. Due to many economic applications this investigation is limited to the two-class problem called binary classification. A typical example is bankruptcy prediction where there are few bankrupted enterprises compared to the sound enterprises in the datasets. In the Polish economy, this is from 0.6% to 4.6% of dataset cardinality, depending on the prediction horizon (one or two years) [Pociecha et al. 2014]. A similar situation takes place in a churn analysis. The number of customers who leave for the competitors is suspected to be a small fraction of the population. When building a model based on data from direct marketing campaigns, the class of customers with a positive response is definitely less numerous. In some applications, e.g. credit card fraud detection, the minority class is hardly 0.1% of the dataset size or even less [Bolton and Hand 2002]. The problem is that when data is unbalanced the classifiers tend to focus on the accurate prediction of the majority class. The reason for this bias depends on the learning method. King and Zeng [2001] show in logistic regression that posterior probabilities of the minority class are underestimated. In turn, classification trees use criteria in recursive partitioning procedure that minimizes overall error regardless of class. For this reason, one usually obtains classifiers with a high accuracy of prediction, that misclassify mainly the objects from the minority class. Note that these objects represent rare events whose explanation the analyst is particularly interested in. Suppose one has data where the minority class is 5%. Having constructed a model with an error of about 0.05 means that one can simply classify everything to the dominating class with the same effect.

The unbalanced learning problem has attracted a great deal of interest recently. In general, solutions for unbalanced data can be summarized in two approaches: modifications at the level of the learning algorithm, or at the data level. The algorithmic changes include cost-sensitive learning and shifting a threshold for posterior probabilities. The changes at data level include resampling and feature selection. Note that resampling usually works as a pre-processing step, but in ensemble learning it can be a part of the learning algorithm. The study of performance of different resampling methods is given e.g. in [Loyola-González et al. 2016]. In order to increase the number of observations from the minority class, Lee [2000] introduced some normal noise to the training set, whereas Chawla et al. [2002] used a distance based algorithm for generating new, artificial objects. Kumar et al. [2014] supported under-sampling with the use of cluster analysis. A comprehensive comparative study can be found in the work of López et al. [2012]. Providing a complete review of literature falls beyond the scope of this paper. An overview of the existing solutions with a wide reference can be found in [Chawla et al. 2004; Haixiang et al. 2017; Longadge et al. 2013; Weiss 2004]. Another view which emphasises ensemble learning was given by Galar et al. [2011]. A separate problem related to unbalanced data is model assessment. An extensive investigation on the estimation of model quality is discussed in [Japkowicz, Shah 2011].

This article focuses on the effectiveness of resampling methods, which can be divided into two groups: random or “intelligent”, i.e. those that use the information contained in the data. This second group is represented by sophisticated methods that introduce additional computational cost to the learning process. The objective is to verify the advantages from such resampling. To answer the question whether resampling is beneficial at all, the authors compare the obtained results with a classification based on the shifted posterior probability threshold. In this second approach a classifier is estimated using the whole unbalanced dataset. Since a family of classifiers may strongly affect the results, this study is limited to the random forests [Breiman 2001] and logistic regression model. This choice is not accidental as random forests were successfully applied in many research areas. This classifier is valued for accuracy of prediction, dealing with various types of variables, robustness on outliers, and automatic feature selection, however the disadvantage of ensembles is the loss of interpretability. For this reason the second classifier taken into consideration is logistic regression. The authors estimate model parameters using the classic criterion of maximum likelihood as well as including the regularization term. Although extensive research on logistic regression for unbalanced data has been carried out, as far as it is known, there is no such study for its regularised version. The authors found it especially interesting because of embedded feature selection. Note that many similar studies have been reported in the literature but the majority of them consider single trees and classical logistic regression, or neural networks and support vector machines. Moreover, the results of these works are sometimes contradictory.

The rest of this paper is organized as follows. Section 2 describes simple measures of model quality that are used in the case of unbalanced data. There are also introduced two types of classifiers that are the focus in the further research. Section 3 discusses balancing techniques. The setup and the results of the research are presented in Section 4, and Section 5 includes the concluding remarks.

2. Classifier assessment in the case of unbalanced data

Assume that a set of multidimensional observations $\{(\boldsymbol{x}_i, y_i): i \in \{1, ..., N\}\}$ is given, where vectors $\boldsymbol{x}_i$ consist of measurements of predictors $\boldsymbol{X} = (X1, ..., X_p)$, and y_i are one of two possible class labels, which will be denoted by $\{0,1\}$. Let us establish that 1 encodes a minority class. The objective is to model dependency $y = f(\boldsymbol{x})$. Function f is called the classification rule or classifier. As a dataset is usually a random sample, one obtains the estimate $\hat{y} = \hat{f}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ where $\boldsymbol{\theta}$ is a vector of model parameters. Then, a classifier is used for the prediction of class for new objects.

This paper considers two types of classifiers. The first one is random forest [Breiman 2001], which represents an ensemble approach to statistical learning. The trees are constructed without pruning and only a small number of randomly picked variables is considered within the nodes. Next the responses of the trees are combined in the final output. The only hyperparameters are the number of trees and the number of picked variables, which are suggested in the source work [Breiman 2001]. This makes this method convenient even for non-advanced data analysts. Note that random forests scale well in high-dimensional domains because only randomly selected features within any node are considered for making a split. Unfortunately, this model works as a black box, so it is not interpretable. The second classifier considered is the logistic regression model; this model is especially popular because of its interpretative possibilities. The parameters of the logit model:

$$ln\frac{Pr(Y=1|x)}{1-Pr(Y=1|x)} = b_0 + b_1x_1 + ... + b_px_p \tag{1}$$

can be estimated in the classic way, i.e. using the maximum likelihood method, or by regularization. In this second approach the penalty component $P(\boldsymbol{b})$ is included to the estimation criterion:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}}(-2\ln L(\boldsymbol{b}) + \lambda \cdot P(\boldsymbol{b})), \tag{2}$$

where $L(\boldsymbol{b})$ is a likelihood function. The penalty causes the shrinking of coefficients to zero. In extreme cases they can be equal to zero, what is tantamount with feature selection. One can consider the penalty in the form of elastic net:

$$P_\alpha(\boldsymbol{b}) = \sum_{j=1}^{p}\left((1-\alpha)b_j^2 + \alpha\left|b_j\right|\right), \tag{3}$$

which was proposed by Zou and Hastie [2005]. It combines the ridge regression and lasso. Parameter λ decides the amount of shrinking and it is usually determined adaptively. Usually several models are estimated for different values of λ and finally one is chosen which minimizes the evaluation function, e.g. information criterion.

An important stage of modelling is an assessment of classifier usefulness. The most popular and frequently used measure of model quality is classification accuracy, i.e. the fraction of correctly classified objects. As previously indicated, this is not a proper measure in the case of unbalanced data. Therefore, several measures were proposed that take into account the type of incorrect classification [Fawcett 2006]. Table 1 consists of the numbers of correct and incorrect classifications with regard to the classes. It also presents the notations commonly used in the literature. Note that the minority class, which is usually the class of interest, is called a positive class while the majority class is a negative.

Table 1. Confusion matrix. Minority class is coded by 1

Observed class	Predicted class	
	0	1
0	*TN*	*FP*
1	*FN*	*TP*
TN (*true negative*) *FP* (*false positive*)	*TP* (*true positive*) *FN* (*false negative*)	

Source: own work.

The simplest evaluation measure which focuses on rare events is the accuracy, calculated for the minority class only:

$$TPR = \frac{TP}{TP + FN}. \tag{4}$$

This is called *True Positive Rate* or *sensitivity*. In turn, the accuracy of classification of the majority class:

$$TNR = \frac{TN}{TN + FP} \tag{5}$$

is called *specificity*. Probably the most popular measure which takes into account both types of accuracies is the so-called *Area Under the Curve* and it is simply the arithmetic mean of *sensitivity* and *specificity*:

$$AUC = 0{,}5 \cdot \left(TPR + TNR\right). \tag{6}$$

The term *AUC* derives from a geometric interpretation in the ROC space [Misztal 2014]. This measure reflects a compromise between the correct classification of both classes. Naturally, in the model selection stage one can consider *sensitivity* as well as the *AUC* measure to choose the final classifier. Note that as a model is usually used for the prediction of future events, the quality measures should be estimated on unseen data, i.e. on test samples independent of the learning stage [Hastie et al. 2009].

3. Data balancing

Balancing is a pre-processing step which is oriented on the preparation of the dataset for a learning algorithm. The idea is that the classes would be approximately equinumerous in the input data, then the classifiers would not be focused on the majority class.

The simplest and algorithmically the fastest method of data balancing is random resampling, which can be achieved in three ways. Under-sampling leaves all objects from the minority class and it randomly eliminates the objects from the majority class. The basic drawback of this approach, indicated in the literature, is that important objects can be potentially removed. Moreover, in the case of small datasets the size of the obtained input training set can be relatively small compared to the number of predictors, which induces a problem with the accuracy of estimation of logistic regression parameters. In an extreme scenario this can be even $N < p$. In turn, in the case of large datasets, under-sampling accelerates the run of the learning algorithm. The second way of random data balancing is over-sampling, which replicates the objects from the minority class. There is no lost information in this approach, but it is commonly believed that the constructed model may then overfit. As an example, let us take a classification tree. The region defined by the rule (conjunction of the conditions on the path from the root to the leaf) may cover a great number of positive objects, that are in fact exact copies of one. In the case of a large dataset, a further increase of size by over-sampling may substantially slow down a learning algorithm. The partial response to the drawbacks of the discussed methods is their combination. The mix of under and over-sampling provides the opportunity of setting the input training set size depending on preferences. Unfortunately, despite many empirical studies reported in the literature (see e.g. [Estabrooks et al. 2004]), there are no clear results on the optimal rate of under and over-sampling in a mixed approach.

The additional possibilities of using random resampling are given by ensemble learning. The typical drawback of under-sampling, i.e. the loss of important information, is overcome in a natural way because a resampling can be repeated for each base model. The implementation of this approach is available directly from standard random forest procedure in R. The method is known as balanced random forest [Chen et al. 2004].

Due to the mentioned drawbacks of random resampling, several more advanced techniques have been proposed. Using information from a dataset, they create new synthetic objects for the minority class, or discard some of the representatives from the majority class. Thus, they are the information-based versions of over or under-sampling. Generally, these methods use distances between objects or distribution estimation in the classes.

The SMOTE algorithm [Chawla et al. 2002] is probably the most popular among these which utilise the nearest neighbours approach. The idea is to create synthetic objects in the neighbourhood of points from the minority class. Namely, for each observation $\boldsymbol{x}$ from a minority class, its k nearest neighbors from this class are determined and then one of them $NN^*(\boldsymbol{x})$ is picked randomly. Next, the coordinate differences between $\boldsymbol{x}$ and $NN^*(\boldsymbol{x})$ are calculated. These differences are multiplied by the random numbers between 0 and 1. The new object is created by adding such a vector of randomly deformed differences to the original vector $\boldsymbol{x}$. Geometrically this looks like a shifting of point $\boldsymbol{x}$ towards nearest neighbor $NN^*(\boldsymbol{x})$. The number k as well as the number of picked neighbors $NN^*(\boldsymbol{x})$ are parameters of the algorithm. The first decides, how much the synthetic observations may be dispersed around the original ones. The second is set according to the required amount of over-sampling. For example, if the size of the minority class is to increase by 300%, only three of the nearest neighbors are picked for each positive object $\boldsymbol{x}$.

The ROSE algorithm [Menardi, Torelli 2014] reflects a distribution-based approach. It generates a new set of synthetic objects in which the classes are equally represented. The size of the new training set is a parameter of the method, thus under-sampling, over-sampling as well as the mixed approach is possible. The algorithm is partially random. It starts from picking an object $(\boldsymbol{x}_i, y_i)$ from the original training set with the same class probabilities. Then new observation $\boldsymbol{x}$ is generated according to density function $f(\mathbf{x}|Y = y_i)$ which is approximated by the kernel estimate:

$$\hat{f}(\mathbf{x}|Y = y_i) = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{H_j}(x - x_i), \tag{7}$$

where j indicates a class, n_j is its size and $\boldsymbol{H}_j$ is a matrix of scale parameters in the chosen class. Menardi and Torelli [2014] applied the Gaussian kernel.

4. Empirical study

The empirical comparison was conducted for two popular and in some sense complementary classifiers. Random forests [Breiman 2001] represent the modern tool with a high accuracy of prediction, which acts like a black box. On the other hand, logistic regression represents an interpretable model with a fast estimation process, and is generally sufficient when the class structure is not very complicated. To obtain results free of influence of irrelevant variables, therefore more comparable with random forests, the study also investigated regularized logistic regression utilising elastic net penalty eq. (3). The setups in this research are as follow. Random forests consist of 200 trees which are built without pruning. The number of variables randomly picked in the nodes is approximately a square root of p (the number of predictors), using R package `randomForest` for this purpose. The logistic regression model is fitted with the use of a coordinate-descent algorithm [Friedman et al. 2008] implemented in R package `glmnet`, which is designed for classic as well as regularized estimation. The alpha parameter in eq. (3) was set as 0.9 to assign more weight to the term with absolute value, which decides about the feature-selection effect. Penalty parameter λ was determined according to the minimal value of the Bayesian information criterion. The notations of these classifiers used in the tables with results are: RF (for random forests), LR (for logistic regression) and RLR (for regularized logistic regression). For resampling methods the study used R packages `ROSE` and `smotefamily`. In the random mixed approach the authors discarded half of the majority class, and then replicated the minority class to achieve the equinumerosity.

Table 2. Datasets summary

Datasets	Number of objects	Number of predictors	Fraction of the minority class
Advertisement	3279	1557	0.140
Bank-marketing	4521	16	0.115
Churn	5000	18	0.141
Polish bankruptcy	4769	64	0.025
Seismic bumps	2584	18	0.066

Source: UCI Machine Learning Repository.

The investigation was carried out on five real datasets from the UCI repository [Dua, Graff 2019]. Each of them has a binary dependent variable. Their characteristics are shown in Table 2. For logistic regression, the qualitative variables were transformed to dummy variables. For a comparison of random forests and logistic regression, one had to eliminate the problem of missing data. Due to their small amount it was decided on the inputation by means in *advertisement* dataset. In turn, in *Polish bankruptcy* data, where there were a lot of NAs, such rows were deleted. The datasets were split on training sets (two-thirds of the original size) and test sets (one-third of the original size).

Table 3 presents the results for five resampling methods. For the first three datasets, the best results for RF are better than the best results for LR and RLR. In the next two datasets, where the fraction of the minority class is only 2.5% and 6.6%, logistic regression can be competitive. Note that in the case of logistic regression, it is not possible to indicate the most favourable resampling method, while random forests almost always achieve their best results after under-sampling. The only exception in RF is *Advertisement* dataset, where under-sampling leads to the second best result. However even here, the second result is better than all those obtained in logistic regressions. Notably simple random resampling was not outperformed by more sophisticated methods. Only the application of SMOTE with regularized logistic regression yielded the best AUC and TPR two times (for *advertisement* and *Polish bankruptcy*), however these results were not as high as in random forests. Regularization brought an improvement in relation to classical logistic regression in three datasets: *advertisement*, *bank marketing* and *Polish bankruptcy*. Note that the especially large difference in AUC and TPR was for *advertisement* dataset. This is a dataset for text categorization where all except three variables are binary with a small fraction of ones. These 1554 variables indicate the presence of words or phrases in the text documents. The application of SMOTE and RLR left only 54 of these variables. Remember that the study still reports the best results from all the resampling methods, because having conducted this experiment there are no explicit recommendations as to which balancing method is best for logistic regression.

Table 3. Comparison of resampling methods

Datasets and resampling	LR		RLR		RF	
Advertisement	AUC	TPR	AUC	TPR	AUC	TPR
Under-sampling	0.749	0.715	0.883	0.795	0.937	0.901
Over-sampling	0.850	0.735	0.915	0.848	0.946	0.914
Mixed	0.763	0.682	0.894	0.821	0.940	0.914
ROSE	0.739	0.927	0.887	0.801	0.544	1
SMOTE	0.846	0.748	0.927	0.874	0.934	0.881
Bank marketing	AUC	TPR	AUC	TPR	AUC	TPR
Under-sampling	0.790	0.754	0.788	0.714	0.843	0.903
Over-sampling	0.804	0.760	0.818	0.789	0.698	0.451
Mixed	0.809	0.777	0.808	0.771	0.767	0.646
ROSE	0.806	0.777	0.809	0.789	0.790	0.720
SMOTE	0.798	0.731	0.788	0.714	0.622	0.269

Churn	AUC	TPR	AUC	TPR	AUC	TPR
Under-sampling	0.796	0.830	0.785	0.808	0.883	0.848
Over-sampling	0.793	0.826	0.796	0.830	0.869	0.746
Mixed	0.776	0.781	0.772	0.781	0.880	0.786
ROSE	0.775	0.790	0.781	0.808	0.852	0.786
SMOTE	0.779	0.754	0.773	0.746	0.880	0.772
Polish bankruptcy	AUC	TPR	AUC	TPR	AUC	TPR
Under-sampling	0.691	0.737	0.709	0.605	0.770	0.842
Over-sampling	0.756	0.868	0.778	0.763	0.524	0.053
Mixed	0.784	0.763	0.781	0.763	0.610	0.237
ROSE	0.761	0.842	0.756	0.789	0.503	1
SMOTE	0.766	0.895	0.819	0.816	0.605	0.237
Seismic bumps	AUC	TPR	AUC	TPR	AUC	TPR
Under-sampling	0.738	0.742	0.744	0.677	0.739	0.758
Over-sampling	0.741	0.726	0.750	0.742	0.536	0.097
Mixed	0.761	0.758	0.755	0.742	0.601	0.258
ROSE	0.746	0.726	0.738	0.742	0.719	0.548
SMOTE	0.743	0.726	0.743	0.726	0.549	0.129

Source: own calculations.

At this stage of research one can ask the question whether to balance a data or not? Under-sampling is evidently the most beneficial technique for random forests. As this decreases the size of the training set, and therefore potentially discards important information, the paper compares this method with learning on a full dataset. In this case the threshold was set for posterior probabilities equal to the fraction of the minority class in a training set. It can be seen that the results (Table 4) are second best or even best for random forest. However, in the case of logistic regression, it only occasionally improves the AUC or TPR.

Table 4. Results obtained using original datasets and shifting a classification threshold

Datasets	LR		RLR		RF	
	AUC	TPR	AUC	TPR	AUC	TPR
Advertisement	0.480	0.252	0.906	0.848	0.949	0.940
Bank marketing	0.802	0.760	0.790	0.737	0.841	0.903
Churn	0.785	0.817	0.793	0.835	0.882	0.866
Polish bankruptcy	0.810	0.842	0.710	0.763	0.731	0.763
Seismic bumps	0.731	0.726	0.749	0.758	0.710	0.712

Source: own calculations.

Table 5. Balanced random forest vs. standard version

Datasets	Under-sampling				Threshold	
	RF		BRF		RF	
	AUC	TPR	AUC	TPR	AUC	TPR
Advertisement	0.937	0.901	0.943	0.921	0.949	0.940
Bank marketing	0.843	0.903	0.834	0.863	0.841	0.903
Churn	0.883	0.848	0.898	0.857	0.882	0.866
Polish bankruptcy	0.739	0.758	0.790	0.868	0.731	0.763
Seismic bumps	0.739	0.758	0.734	0.774	0.710	0.712

Source: own calculations.

Up to now the authors were combining random forests with data balancing in a pre-processing step, i.e. resampling was performed outside of the learning algorithm, thus, not using entirely the possibilities of ensemble learning. In balanced random forest BRF [Chen et al. 2004], under-sampling is performed for each base model. Usually this gives a slight improvement in the results of AUC and TPR (Table 5), but the differences are unexpectedly low except *Polish bankruptcy* data, where AUC increased by about 0.05, and TPR by about 0.11. Yet, the thresholding of posterior probabilities can be even competitive when a fraction of the minority class is more than 10%. In the case of small fractions (*Polish bankruptcy* and *seismic bumps* datasets), resampling was more advantageous.

5. Conclusion

Resampling methods are the most popular remedy for the class unbalance problem [Haixiang et al. 2017]. In the case of random forests, random under-sampling returned the best results unambiguously. This connection was usually substantially superior to any over-sampling method, random or information based. One can say that random forests do not tolerate artificially replicated information. The predominance of under-sampling, a technique that reduces the size of the training set, was a somewhat surprising result, because the trees require a large sample. In fact, it is beneficial in the case of processing large datasets. However, a high unbalance or low size of the original dataset in connection with under-sampling may lead to very small training sets. This could be problematic and the authors set it as a direction for future work. Moreover, logistic regression may be competitive when a fraction of the minority class is low. Unfortunately, none of the investigated resampling method has obtained superiority for this model. Thus, the choice of resampling would have to be performed with the use of a validation set. This would induce an additional split of the original data and would decrease the size of the training sample. It is noteworthy that learning on whole datasets and classification according to the shifted threshold frequently returns at least comparable results when the fraction of the minority class is not too low. Thus, for smaller datasets one can run random forests without under-resampling and shift a classification threshold. In the case of regularized logistic regression, this approach could be a solution to the problem with a choice of resampling method, unfortunately encountering a problem at the lambda selection stage, bearing in mind that lambda is chosen so as to minimize the BIC criterion. The component of BIC is deviance, which reflects a goodness of fit. When data is unbalanced, a deviance does not distinguish the kind of incorrect classifications. Package `glmnet` in R implements cross-validation where deviance is also minimized. Therefore the selection of lambda so that it maximizes AUC is a second direction of the authors future work.

References

Bolton R.J., Hand D.J., 2002, *Statistical fraud detection*, Statistical Science, vol. 17, no. 3, 235-255.

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*, University of California,: School of Information and Computer Science, Irvine, CA http://archive.ics.uci.edu/ml

Estabrooks A., Jo T., Japkowicz N., 2004, *A multiple resampling method for learning from imbalanced data sets*, Computational Intelligence, 20(1), 18-36.

Fawcett T., 2006, *An introduction to ROC analysis*, Pattern Recognition Letters, 27, 861-874.

Friedman J., Hastie T., Tibshirani R., 2008, *Regularization paths for generalized linear models via coordinate descent*, Technical report, Stanford University.

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.

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.

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

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.

Kumar N.S., Rao K.N., Govardhan A., Reddy K.S. & Mahmood A.M., 2014, *Undersampled k-means approach for handling imbalanced distributed data*, Progress in Artificial Intelligence, 3(1), 29-38.

Lee S., 2000, Noisy replication in skewed binary classification, Computational Statistics and Data Analysis, 34, 165-191.

Longadge R., Dongre S.S., Malik L., 2013, *Class imbalance problem in data mining: review,* International Journal of Computer Science and Network, vol. 2, issue 1, 83-87.

Loyola-González O., Martínez-Trinidad J. F., Carrasco-Ochoa J.A., García-Borroto M., 2016, *Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases*, Neurocomputing, 175, 935-947.

López V., Fernández A., Moreno-Torres J. G., & Herrera F., 2012, *Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics*, Expert Systems with Applications, 39(7), 6585-6608.

Menardi G., Torelli N., 2014, *Training and assessing classification rules with imbalanced data,* Data Mining and Knowledge Discovery, 28, 92-122.

Misztal M., 2014, *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 328, Taksonomia 23, Klasyfikacja i analiza danych – teoria i zastosowania, 156-166.

Pociecha J., Pawełek B., Baryła M., Augustyn S., 2014, *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*, Fundacja Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Weiss G., 2004, *Mining with rarity: A unifying framework*, SIGKDD Explorations, 6(1), 7-19.

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.

OCENA METOD REPRÓBKOWANIA W PROBLEMIE ZBIORÓW NIEZBILANSOWANYCH

Streszczenie: Celem wielu praktycznych zastosowań modeli dyskryminacyjnych jest przewidywanie zdarzeń rzadkich. Zbiory uczące są wówczas niezbilansowane. W tym przypadku klasyfikatory mają tendencję do poprawnego klasyfikowania obiektów klasy większościowej i jednocześnie błędnie klasyfikują wiele obiektów klasy mniejszościowej, która jest przedmiotem szczególnego zainteresowania. W celu rozwiązania tego problemu zaproponowano wiele technik, które bilansują dane lub modyfikują algorytmy uczące. Celem artykułu jest porównanie prostych, losowych metod bilansowania z bardziej wyrafinowanymi, które pojawiły się w literaturze. Dodatkowo postawiono pytanie, czy konkurencyjnym podejściem nie jest budowa modelu na oryginalnym zbiorze danych i przesunięcie progu klasyfikacji. Badanie przedstawiono z perspektywy regularyzowanej regresji logistycznej i lasów losowych. Wyniki pokazują, że kombinacja metody under-sampling z lasami losowymi wykazuje przewagę nad innymi technikami, podczas gdy regresja logistyczna może być konkurencyjna w przypadku silnego niezbilansowania.

Słowa kluczowe: klasy niezbilansowane, repróbkowanie, regularyzowana regresja logistyczna, lasy losowe.