UDC 001.81:004.8
Original Scientific Paper

Two Stage Comparison of Classifier Performances for Highly Imbalanced Datasets

Goran Oreški *goran.oreski@gmail.com*
GO Studio Ltd.
Unska 54, 44324 Jasenovac, Croatia

Stjepan Oreški *stjepan.oreski@kaba.hr*
Bank of Karlovac
I. G. Kovačića 1, 47000 Karlovac, Croatia

Abstract

During the process of knowledge discovery in data, imbalanced learning data often emerges and presents a significant challenge for data mining methods. In this paper, we investigate the influence of class imbalanced data on the classification results of artificial intelligence methods, i.e. neural networks and support vector machine, and on the classification results of classical classification methods represented by RIPPER and the Naïve Bayes classifier. All experiments are conducted on 30 different imbalanced datasets obtained from KEEL (Knowledge Extraction based on Evolutionary Learning) repository. With the purpose of measuring the quality of classification, the accuracy and the area under ROC curve (AUC) measures are used. The results of the research indicate that the neural network and support vector machine show improvement of the AUC measure when applied to balanced data, but at the same time, they show the deterioration of results from the aspect of classification accuracy. RIPPER results are also similar, but the changes are of a smaller magnitude, while the results of the Naïve Bayes classifier show overall deterioration of results on balanced distributions. The number of instances in the presented highly imbalanced datasets has significant additional impact on the classification performances of the SVM classifier. The results have shown the potential of the SVM classifier for the ensemble creation on imbalanced datasets.

Keywords: imbalanced data, classification algorithm, re-sampling technique, dataset cardinality, reduction of class imbalance

1. Introduction

The ongoing trend of exponential growth of available data makes the process of knowledge discovery in data (KDD) even more important. Thereby, the most challenging problems are in the field of classification. Real-world classification problems have resulted in the vast number of cases where classification learning is additionally difficult because of imbalanced data sets. Such cases can be found in medicine, financial industry, chemistry, engineering and other real-world domains where machine learning is used for data classification problems.

The imbalance of data in this paper refers to inter-class imbalance, i.e. the case when some classes have much more examples than others. The imbalance is expressed through the imbalance ratio (IR), which is defined as the ratio of the number of cases in the majority class according to the number of examples in the minority class. By convention, in imbalanced data sets, we call the classes having more examples majority classes and the ones having fewer examples minority classes. Also, the class label of the minority class is positive, and the class label of the majority class is negative [9]. The fundamental issue with the imbalanced learning problem is the ability of imbalanced data to significantly compromise the performance of

[1] A preliminary version of this study was presented at the 25th Central European Conference on Information and Intelligent Systems, CECIIS 2014, held in Varaždin, Croatia, September 17-19, 2014 [19].

most advanced learning algorithms. The most advanced algorithms assume or expect balanced class distributions or equal misclassification costs. Therefore, when presented with complex imbalanced data sets, these algorithms fail to properly represent the distributive characteristics of the data and resultantly provide unfavorable accuracies across the classes of the data [10].
In recent years there have been many scientific papers that address this topic. Most of the papers are focused on finding the best classification algorithm for a certain dataset or datasets [3],[17], as well as on proposing new techniques for data re-sampling [4],[9].

The main goal of the study presented in this paper is to explore the key characteristics of certain classification algorithms, i.e., the key characteristics of strategies on which classification algorithms are based, **with regard to imbalanced datasets**. The characteristics of selected algorithms are considered on original datasets, that is original distributions, and on balanced datasets. An additional goal is to explore the key characteristics of the classification algorithms with regard to the number of instances in these datasets.

This paper is organized as follows. Section 2 describes the problem of imbalanced data and their influence on classification algorithms and reviews the literature related to the problem. In Section 3 we very briefly describe the fundamental characteristics of each selected classification algorithm and the SMOTE technique. Section 4 describes the experimental design. In Section 5 we provide empirical results with a discussion. Section 6 concludes this paper and gives some guidelines for future work.

2. Problem statement and literature review

During the learning process, sophisticated classification algorithms are guided towards maximizing the accuracy of the classification prediction. In the real world there are cases in which maximal accuracy is not the goal of classification, therefore such algorithms, without the application of some additional preprocessing techniques, are not necessarily the best choice.

The focus of this research is to analyze the key characteristics of certain classification algorithms with regard to imbalanced datasets. This is performed thru: (1) the impact analysis of the additional preprocessing technique application on classification algorithms' performances and (2) the impact analysis of the number of instances in a dataset on certain classification algorithms' performances.

The literature in the field of class imbalance is vast. One of the first studies which brought together the previous research work is the Japkowicz paper [11]. It concluded that while a standard multilayer perceptron neural network is not sensitive to the class imbalance problem when applied to linearly separable domains, its sensitivity increases with the complexity of the domain.

The most common topics of the research are; the creation of a new technique for data balancing [4],[9], analysis of the relationship between class imbalance and the cost of miss-classification [5], the impact of imbalanced and noisy data to classifier performances [8],[23], the research of different evaluation measures in class imbalance conditions [20], finding the best strategies for establishing the optimal balance ratio in imbalanced data [6].
According to the topic of this research, in the next section we provide a short description of the selected algorithms, whose performances are the subject of the study.

3. Methodological backgrounds

According to the primary goal of the paper, we have selected four algorithms to investigate to which extent they perform on imbalanced data sets. The following algorithms were selected for experiments: back propagation neural network, linear support vector machine, ripper and naïve Bayes. In order to achieve the purpose of this study, in this section we will briefly describe the algorithms used in the research. Additionally, we provide a short description of the SMOTE technique, used for distribution balancing of datasets.

3.1. Neural network

Neural networks (NN) are part of the computational and artificial intelligence field and therefore can be classified as an artificial intelligence method. There are many different kinds of neural networks and neural network algorithms. The neural network algorithm used in the experiments is the most representative and popular algorithm called back-propagation. Multilayer feed-forward network is the type of neural network on which the back-propagation algorithm performs [18]. This algorithm is a variation of the gradient descent algorithm to find a minimum of an error function in the weight space [15]. As stated earlier, NN tend to have best performance on balanced class distributions, their performance on imbalanced datasets is a part of this research.

3.2. Support vector machine

Support vector machine (SVM) belongs to the same field as the neural networks. In their simplest form, SVMs are based on hyperplanes that separate the training data by a maximal margin. All vectors lying on one side of the hyperplane are labeled as -1, and all vectors lying on the other side are labeled as 1. The training instances that lie closest to the hyperplane are called support vectors [22]. This artificial intelligence method has been very successful in application areas ranging from image retrieval, handwriting recognition to text classification [1]. However, when faced with imbalanced datasets where the number of negative instances by far outnumbers the positive instances, the performance of SVM drops significantly [24].

3.3. RIPPER (Repeated Incremental Pruning to Produce Error Reduction)

As an example of a classical algorithmic approach to solving the class imbalance problem, the simple rule induction learning algorithm, RIPPER is used. The RIPPER algorithm is a rule induction system which makes use of a divide and conquers a strategy to create a series of rules which describe a specific class. It builds a series of rules for each class, even for very rare classes. It has been shown its particular use, especially with the highly skewed noisy datasets containing many dimensions [2].

3.4. Naïve Bayes

Probabilistic classifiers and, in particular, the naïve Bayes classifier, are among the most popular classifiers in the machine learning community and they are used increasingly in many applications [13]. The naive Bayes classifier greatly simplifies learning by assuming that features are independent of a given class. Bayesian classifiers assign the most likely class to a given example described by its feature vector. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers [21].

3.5. SMOTE

In the SMOTE (Synthetic Minority Over-sampling Technique) technique, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors [4]. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. The default implementation uses five nearest neighbors. This approach effectively forces the decision region of the minority class to become more general [4].

4. Research design

This section describes the research design that has been proposed to deal with questions of interest. In doing so, firstly, two different procedures used in this research are described, and after that measures for evaluating the results are presented.

As previously mentioned, rather than finding the best classification method, this study highlights the capabilities of learning strategies presented here according to their efficiency to address classification with imbalanced data, with and without using a re-sampling technique. An additional analysis was performed to determine the impact of the number of instances vis-a-vis the classifier performances. Four learning algorithms are selected, all from the RapidMiner machine learning toolkit, Version 5.3 on Intel Core i3 CPU 2.13 GHz, 4GB of RAM. These learning algorithms are; back propagation neural network (NN), linear support vector machine (SVM), Ripper (RIP, implementation as Weka:W-JRip), and naive Bayes (NB). They represent a diverse set of well-known learning strategies as are considered in the Methodological background section. We used the default parameter values in each case for each algorithm, because our main aim was to highlight the differences in their basic performance, measured with and without the SMOTE re-sampling technique, and not to find the best classifier.

4.1. Research procedure description

Initially, 30 different imbalanced datasets are selected from the KEEL repository. Each original dataset is presented as the input of four selected learning algorithms. The 10-fold cross-validation technique is used in order to create and validate the performance of the models. Second procedure, with the SMOTE technique included, was different. In this procedure a preprocessing step is added. All datasets are re-sampled, i.e. balanced with the SMOTE technique. Balanced datasets are taken as the input of the four selected learning algorithms. So created models are validated against the original datasets. Validation with original datasets, according to Brennan [2], is the best method of validation in such circumstances.

After analyzing all the results obtained in the first experiment, in the second experiment an additional result analysis of all datasets in terms of the number of instances in the original datasets was performed. Here the datasets were divided in two groups. Each group contains approximately the same number of sets. All the results of the classification and validation are recorded in the form of the confusion matrix. From these results, two performance measures are calculated; accuracy and AUC.

Accuracy is a good, integrated evaluation measure if the balance between positive and negative examples exists. When used to evaluate the performance of a learner for imbalanced data sets, accuracy is generally better suitable to evaluate the majority class and behaves poorly to the minority class. Accordingly, if the dataset is extremely imbalanced, even when the classifier classifies all the majority examples correctly and misclassifies all the minority examples, the accuracy of the learner is still high because there are many more majority examples than minority examples. Under this circumstance, accuracy cannot evaluate prediction for the minority class reliably. Thus, more reasonable evaluation metrics are needed. In such circumstance, the Area Under the ROC Curve (AUC) is accepted as a traditional performance metric. The AUC is the classifier quality measure independent of the imbalance ratio, i.e., not biased in favor of any class. The AUC is a good way to get a score for the general performance of a classifier and to compare it with that of another classifier. This is particularly true in the case of imbalanced data in which accuracy is strongly biased toward the dominant class [12]. Accordingly, in our research the AUC is used as the second measure.

4.2. Statistical comparisons

The first experiment research results are verified by statistical tests. The results of each dataset are tested before and after balancing. From the statistical point of view, we are comparing the performance of two classifiers on a single domain every time. Testing was performed by a paired t-test, one of the most widely used statistical significance measures currently adopted in the context of classifier evaluation [12]. Additional statistical testing was done with a nonparametric alternative that is convenient for comparing two classifiers on a single domain; the Wilcoxon matched pairs signed ranks test [7]. In order to reduce the likelihood of the type I error, tests were made with the significance a=0.01. In the second experiment, to determine whether the means of two groups are equal to each other, two-sample t-test assuming unequal variances is used. Two-sample t-test assuming unequal variances, known as Welch's t-test, adjusts the number of degrees of freedom when the variances are thought not to be equal to each other [14]. Tests were conducted with the level of significance a=0.05.

The research results are finally presented in tables and line diagrams.

5. Results and discussion

As was mentioned in the research procedure description, the research was performed in two stages, i.e., two experiments.

5.1. Experiment 1: General comparison of classifier performances for highly imbalanced datasets

Experiment 1 was conducted on 30 different datasets, obtained from the KEEL (Knowledge Extraction based on Evolutionary Learning) repository, with a wide variety of class distributions and with a different number of observations in datasets. In these datasets, the imbalance ratio goes from 9:1 to 41:1, and the number of observations goes from 92 to 1829. List of datasets with their number of instances and the imbalance ratio is given in Table 3.

In Table 1, the accuracy of all four classifiers on thirty class imbalance datasets is shown. In the column named “Original” the accuracy of the original dataset is shown, while in the column “SMOTE” the accuracy of the balanced dataset is shown. The table shows that all four classifiers have better average accuracy scores on original datasets. For each classifier, to compare average accuracy scores before and after data balancing, two-tailed paired t-tests were applied. The minimal number of observations in selected datasets is enough for the application of this statistic. In Table 1 the corresponding p-values are shown. The null hypothesis is that there is no statistically significant difference between the average accuracy before and after data balancing. According to t-tests, we can reject the null hypothesis for the NN, RIP and NB classifiers because the calculated p-values are smaller than the chosen level of significance a=0.01. T-test is not applicable to SVM, because the pairing was not significantly effective, i.e., the differences between paired values are not consistent [16]. An additional statistical test was done with the nonparametric Wilcoxon matched pairs signed ranks test. This test does not require the same assumptions as t-test. According to p-values for the two-tailed Wilcoxon’s matched-pairs signed rank test, for the significance level of a = 0.01, the median difference between all the classifiers before and after balancing, is significant.

In Table 2 we report the AUC obtained by the selected classifiers before and after the datasets balancing. Table 2 shows those classifiers: NN, SVM and RIP have better average AUC scores on balanced (SMOTE) datasets, while the NB classifier has a better average AUC score on original datasets.

Dataset	NN		SVM		RIP		NB	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
cleveland-0vs4	0,9474	0,9595	0,9536	0,6705	0,8958	0,9711	0,9301	0,9191
ecoli-01vs235	0,9754	0,9549	0,9508	0,8852	0,9672	0,9467	0,9098	0,9672
ecoli-01vs5	0,9792	0,9250	0,9667	0,9375	0,9792	0,9583	0,9792	0,8042
ecoli-0137vs26	0,9929	0,9146	0,9751	0,8221	0,9893	0,9680	0,9502	0,8007
ecoli-0147vs2356	0,9792	0,9137	0,9167	0,9137	0,9762	0,9613	0,9315	0,9137
ecoli-0147vs56	0,9730	0,9669	0,9701	0,9066	0,9458	0,9639	0,9580	0,9337
ecoli-0347vs56	0,9767	0,9300	0,9222	0,9027	0,9728	0,9611	0,7588	0,3891
ecoli4	0,9911	0,9613	0,9405	0,9435	0,9881	0,9673	0,9375	0,8542
glass-0146vs2	0,9174	0,6976	0,9174	0,3122	0,8974	0,8000	0,4431	0,4146
glass-015vs2	0,9012	0,4767	0,9012	0,1802	0,9302	0,8953	0,4419	0,4070
glass-016vs2	0,9115	0,7708	0,9115	0,2708	0,9427	0,8385	0,4219	0,3906
glass-016vs5	0,9565	0,9728	0,9511	0,8098	0,9946	0,9728	0,9783	0,8641
glass-04vs5	0,9565	0,9239	0,9022	0,8913	0,9891	0,9891	0,9891	0,4457
glass-06vs5	0,9537	0,9907	0,9167	0,7500	0,9907	0,9537	0,9907	0,7870
glass2	0,9206	0,8645	0,9206	0,3224	0,9439	0,9252	0,4579	0,4533
glass4	0,9439	0,9579	0,9393	0,8738	0,9813	0,9626	0,9019	0,8505
led-02456789vs1	0,9549	0,9300	0,6187	0,8533	0,9617	0,9549	0,8985	0,8262
page-blocks13vs4	0,9576	0,9725	0,9661	0,9343	0,9957	0,9873	0,9386	0,9534
shuttle-c0vsc4	0,9995	1,0000	1,0000	1,0000	1,0000	1,0000	0,9989	0,9978
yeast-0256vs3789	0,9313	0,8825	0,9084	0,8705	0,9502	0,9133	0,9163	0,9203
yeast-02579vs368	0,9641	0,9333	0,9691	0,9293	0,9561	0,9622	0,8884	0,7590
yeast-0359vs78	0,9091	0,8399	0,9170	0,7273	0,9289	0,8439	0,5652	0,3439
yeast-05679vs4	0,9375	0,8333	0,9034	0,8182	0,9527	0,8674	0,5473	0,2879
yeast-1vs7	0,9346	0,7908	0,9346	0,7691	0,9651	0,9172	0,5163	0,3203
yeast-1458vs7	0,9567	0,5758	0,9567	0,6335	0,9567	0,8874	0,2063	0,1573
yeast-2vs4	0,9689	0,9436	0,9339	0,9339	0,9747	0,9533	0,8677	0,4844
yeast-2vs8	0,9793	0,9772	0,9793	0,9772	0,9834	0,9772	0,9647	0,4938
yeast4	0,9670	0,7615	0,9656	0,8592	0,9737	0,9602	0,7460	0,3194
yeast5	0,9805	0,9501	0,9704	0,9259	0,9892	0,9939	0,8996	0,8625
yeast6	0,9805	0,9137	0,9764	0,8895	0,9899	0,9832	0,6442	0,4292
Average	**0,9566**	**0,8828**	**0,9318**	**0,7838**	**0,9654**	**0,9412**	**0,7859**	**0,6450**
Paired t-test (Two-tailed p value[a])	0,001		NA		0,001		0,000	
Wilcoxon matched-pairs signed rank test (Two-tailed p value[a])	0,000		0,000		0,000		0,000	

[a] *level of significance a=0.01.*
Note: A "NA" means not applicable test.
*Notes: An "*Original*" indicates the original dataset while a "*SMOTE*" indicates a balanced dataset.*

Table 1. Accuracy of classifiers on selected imbalanced datasets before and after balancing

Applied statistics, two-tailed paired t-test and the two-tailed Wilcoxon's matched-pairs signed rank test, show that the AUC differences within the NN, SVM and RIP classifiers are statistically significant before and after the datasets balancing. Only the NB classifier has

better average AUC score on original datasets than "SMOTED", but this difference is not statistically significant.

Dataset	NN		SVM		RIP		NB	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
cleveland-0vs4	0,7952	0,9428	0,7630	0,6452	0,6611	0,9137	0,8918	0,8856
ecoli-01vs235	0,8936	0,9564	0,7500	0,8621	0,9447	0,9333	0,5602	0,9261
ecoli-01vs5	0,9205	0,9364	0,8000	0,9205	0,9205	0,9318	0,8977	0,8705
ecoli-0137vs26	0,8571	0,9562	0,5000	0,9088	0,8553	0,9836	0,9745	0,8978
ecoli-0147vs2356	0,8949	0,5000	0,5172	0,5000	0,9089	0,8539	0,6347	0,5000
ecoli-0147vs56	0,8935	0,9270	0,8367	0,9128	0,7502	0,9437	0,7567	0,9274
ecoli-0347vs56	0,8978	0,9434	0,6000	0,9104	0,9314	0,9784	0,8664	0,6616
ecoli4	0,9484	0,9794	0,5000	0,9699	0,9468	0,9592	0,9668	0,9225
glass-0146vs2	0,5000	0,8351	0,5000	0,6250	0,5161	0,891	0,5898	0,6541
glass-015vs2	0,5000	0,7097	0,5000	0,5452	0,6732	0,8896	0,6380	0,5924
glass-016vs2	0,5000	0,7681	0,5000	0,6000	0,7561	0,8318	0,6297	0,6126
glass-016vs5	0,5556	0,9857	0,5000	0,6365	0,9971	0,9857	0,9886	0,9286
glass-04vs5	0,7778	0,9578	0,5000	0,7416	0,9940	0,9940	0,9940	0,6928
glass-06vs5	0,7222	0,9949	0,5000	0,6616	0,9949	0,9747	0,9949	0,8838
glass2	0,5000	0,8189	0,5000	0,6320	0,6471	0,8788	0,6518	0,6762
glass4	0,6104	0,9776	0,5000	0,8249	0,9541	0,9441	0,5880	0,7405
led-02456789vs1	0,8648	0,9495	0,7796	0,8954	0,8931	0,9262	0,8709	0,8560
page-blocks13vs4	0,7098	0,9854	0,7310	0,7476	0,9810	0,9932	0,7666	0,9418
shuttle-c0vsc4	0,9959	1,0000	1,0000	1,0000	1,0000	1,0000	0,9994	0,9951
yeast-0256vs3789	0,6965	0,8179	0,5533	0,8067	0,8015	0,8755	0,6747	0,7849
yeast-02579vs368	0,8722	0,9180	0,8839	0,9113	0,8677	0,9655	0,8841	0,8213
yeast-0359vs78	0,5934	0,7776	0,6067	0,7418	0,6934	0,8243	0,6875	0,6004
yeast-05679vs4	0,7378	0,8377	0,5000	0,7943	0,8162	0,8916	0,7057	0,5708
yeast-1vs7	0,5775	0,7796	0,5000	0,7679	0,7488	0,8782	0,7103	0,6364
yeast-1458vs7	0,5000	0,7146	0,5000	0,6493	0,5000	0,8298	0,5852	0,5596
yeast-2vs4	0,8868	0,9600	0,6667	0,9022	0,8900	0,9392	0,8829	0,6877
yeast-2vs8	0,7739	0,8446	0,7739	0,8446	0,8239	0,9163	0,8142	0,6881
yeast4	0,5291	0,8576	0,5000	0,8609	0,7122	0,847	0,8117	0,6381
yeast5	0,8027	0,9743	0,5000	0,9618	0,9724	0,9859	0,9483	0,9292
yeast6	0,6694	0,9001	0,5000	0,8876	0,8275	0,8798	0,8178	0,7077
Average	**0,7326**	**0,8835**	**0,6087**	**0,7889**	**0,8326**	**0,9213**	**0,7928**	**0,7596**
Paired t-test (Two-tailed p value[a])	0,000		0,000		0,000		0,183	
Wilcoxon matched-pairs signed rank test (Two-tailed p value[a])	0,000		0,000		0,000		0,058	

[a] *level of significance a=0.01.*

Table 2. AUC of classifiers on selected imbalanced datasets before and after balancing

Finally, in Figure 1 and 2, we directly compare the average accuracy and the average AUC obtained by the selected classifiers.

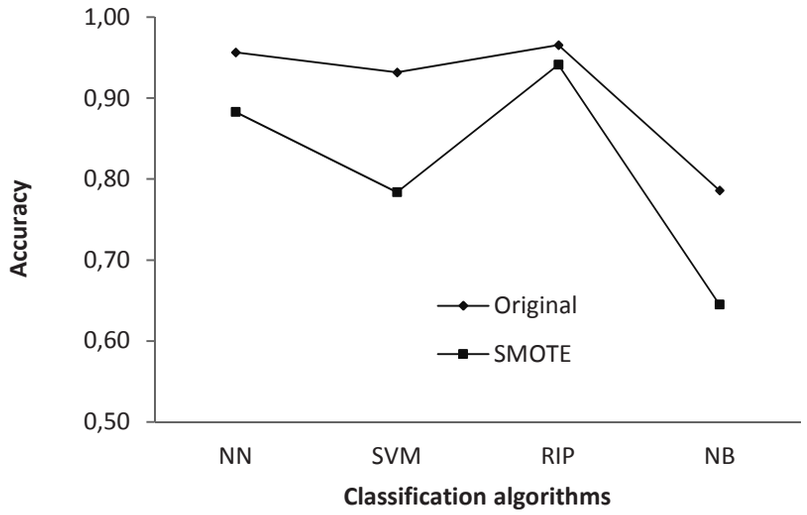

Figure 1. Comparison of the average accuracy of classifiers on original and balanced datasets

The results from this stage of the empirical study indicate that the Ripper classifier is able to cope comparatively well with pronounced class imbalances. At this classifier, the balancing of the sets has a negative impact on classification accuracy, but at the same time has a stronger positive effect on the AUC measure. Very similar characteristics can be attributed to the NN classifier.

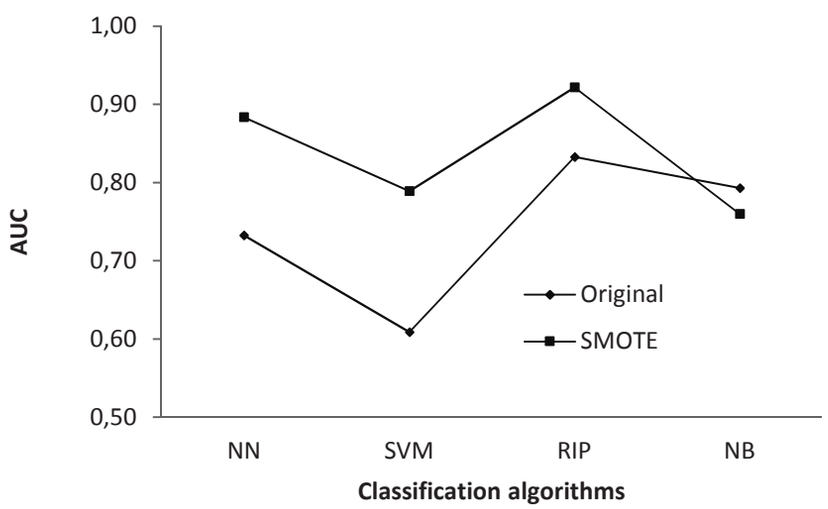

Figure 2. Comparison of the average AUC of classifiers on original and balanced datasets

We also found that, when faced with a large class imbalance, the linear support vector machine algorithm performs significantly worse after balancing training datasets, according to accuracy measure. At the same time, according to the AUC measure, without the balancing the linear support vector machine algorithm performs the poorest. This finding is consistent with the findings of Brown and Mues. They concluded that the use of a linear kernel SVM would not be beneficial in the scoring of data sets where a very large class imbalance exists [3].

Finally, the results from this stage of the research show that imbalanced data have a significant negative influence on the AUC measure at the neural network classifier and, even more, at the linear support vector machine. The same methods show improvement of the AUC measure when applied to balanced data, but at the same time, they show the deterioration of

results from the aspect of classification accuracy. The performances of the Ripper classifier are positively correlated with NN and SVM, but the changes are of a smaller magnitude, while the results of the Naïve Bayes classifier show overall deterioration of results on balanced distributions.

5.2. Experiment 2: An in-depth comparison of classifier performances for highly imbalanced datasets

Until now we analyzed the differences in the performances of the observed classifiers, before and after the balancing of datasets. In the second stage of this research we will explore whether the cardinality of the datasets has an impact on the performances of the observed classifiers. Therefore, we have divided an initial set of 30 observed datasets into two groups. In the first group (Group1) datasets whose number of instances in the set was less than 350 were entered. Group1 consists of 16 datasets. In the second group (Group2) datasets whose cardinality were higher than 350 were entered. Group2 consists of 14 datasets (Table 3). The average number of instances in Group1 was 223.75, while the average number of instances in Group2 was 884.71. According to the cardinality, datasets in Group2 were on average 3.95 times larger than the datasets in Group1.

As in Experiment 1, in this experiment we examine the behavior of the classifier, i.e., we examine the classifier performance for the original datasets and the balanced datasets with the SMOTE technique, but this time for Group1 and Group2 separately.

Group 1			Group 2		
Dataset	Number of instances	Imbalance ratio	Dataset	Number of instances	Imbalance ratio
glass-04vs5	92	9.22	led-02456789vs1	443	10.97
glass-06vs5	108	11.00	yeast-1vs7	459	14.30
glass-015vs2	172	9.12	page-blocks-13vs4	472	15.86
cleveland-0vs4	173	12.62	yeast-2vs8	482	23.10
glass-016vs5	184	19.44	yeast-0359vs78	506	9.12
glass-016vs2	192	10.29	yeast-2vs4	514	9.08
glass-0146vs2	205	11.06	yeast-05679vs4	528	9.35
glass2	214	11.59	yeast-1458vs7	693	22.1
glass4	214	15.47	yeast-0256vs3789	1004	9.14
ecoli-01vs5	240	11.00	yeast-02579vs368	1004	9.14
ecoli-01vs235	244	9.17	yeast4	1484	28.10
ecoli-0347vs56	257	9.28	yeast5	1484	32.73
ecoli-0137vs26	281	39.14	yeast6	1484	41.40
ecoli-0147vs56	332	12.28	shuttle-c0vsc4	1829	13.87
ecoli-0147vs2356	336	10.59	-		
ecoli4	336	15.80	-		

Table 3. Division of datasets into two groups

The measures of performance remain the same; accuracy and AUC. In Table 4 are shown the average classifier accuracies for original and balanced datasets of Group1 and Group2. In the row named “Original” the average accuracy of the original datasets is shown, while in the “SMOTE” row the average accuracy of the balanced datasets is shown for Group1 and Group2 separately. In original datasets, regardless of the set cardinality, the data indicate that all classifiers have almost equal average accuracy scores in both groups, except the NB classifier.

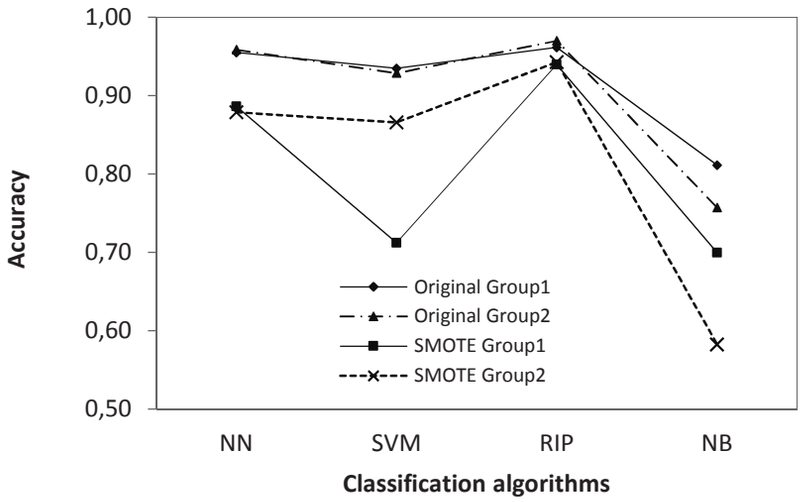

Figure 3. Comparison of the average accuracy of classifiers on original and balanced datasets by groups

The average accuracy in the balanced datasets is slightly different. The biggest difference is shown by the SVM classifier while the NN and RIP classifiers give very stable results. It is very well depicted in Fig. 3.

In Table 5 the average AUC of classifiers by Group1 and Group2 for original and balanced datasets are shown. The data indicate that all classifiers, except the SVM, have very similar average AUC scores on original datasets regardless of the set cardinality. In the balanced datasets, the average AUC results also show similar characteristics. Again, the biggest difference is shown by the SVM classifier, while the NN and RIP classifiers give very uniform results. In Figure 4 we directly compare the average AUC obtained by the selected classifiers on the original and SMOTE sets of Group1 and Group2.

Dataset	NN		SVM		RIP		NB	
	Group1	Group2	Group1	Group2	Group1	Group2	Group1	Group2
Original	0,9548	0,9587	0,9347	0,9285	0,9615	0,9699	0,8112	0,7570
SMOTE	0,8863	0,8789	**0,7120**	**0,8658**	0,9397	0,9430	0,6997	0,5825

Table 4. Average accuracy of classifiers on original and SMOTE datasets shown by groups

Dataset	NN		SVM		RIP		NB	
	Group1	Group2	Group1	Group2	Group1	Group2	Group1	Group2
Original	0,7354	0,7293	0,5792	0,6425	0,8407	0,8234	0,7890	0,7971
SMOTE	0,8868	0,8798	**0,7435** [a]	**0,8408** [b]	0,9305	0,9109	0,7733	0,7441

[a] *In statistical tests shown as SVMGroup1$AUC_SMOTE.*
[b] *In statistical tests shown as SVMGroup2$AUC_SMOTE.*

Table 5. Average AUC of classifiers on original and SMOTE datasets shown by groups

Figure 4 shows that the performance of SVM classifiers is the most sensitive to the cardinality of sets. These results will be additionally statistically analyzed.

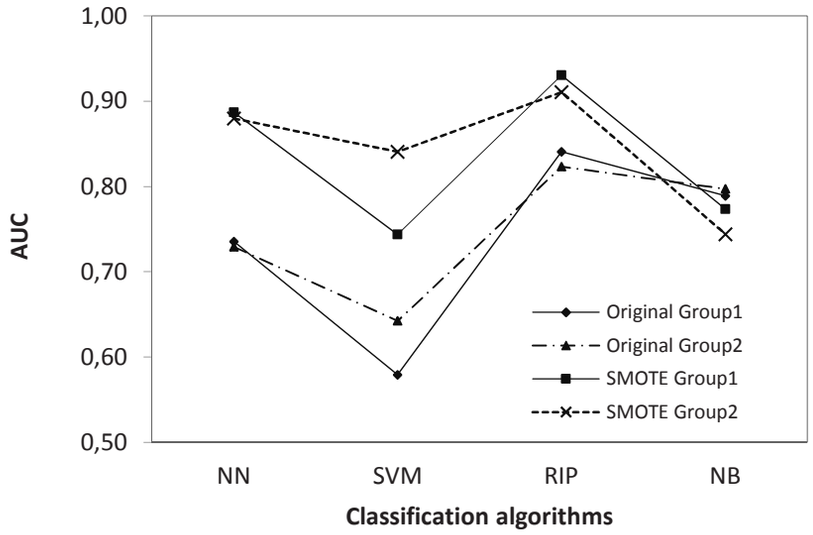

Figure 4. Comparison of the average AUC of classifiers on original and balanced datasets by groups

As was described in the statistical comparison section, to determine whether the measured performance means of Group1 and Group2 are equal to each other, Welch's t-test is used. Tests were conducted with the significance level a=0.05. The assumption for the test is that both groups are sampled from normal distributions. The null hypothesis is that the two means are equal, and the alternative is that they are not. Before we can use the test, we need to perform a normality test to check if groups are sampled from normal distributions. Tests will be performed by the Pearson chi-square normality test and the Shapiro-Wilk normality test.

Only the results of the SVM classifier for the AUC measure on the SMOTE datasets meet normality test assumptions for the application of two-tailed Welch's corrected unpaired t-test. Data analysis is performed by the R version 3.0.3 statistical software. The results of these tests are as follows:

```
Pearson chi-square normality test
data:  SVMGroup1$AUC_SMOTE
P = 6.75, p-value = 0.1497
data:  SVMGroup2$AUC_SMOTE
P = 2.2857, p-value = 0.5153

Shapiro-Wilk normality test
data:  SVMGroup1$AUC_SMOTE
W = 0.9021, p-value = 0.08687
data:  SVMGroup2$AUC_SMOTE
W = 0.9809, p-value = 0.9796
```

As all p-values are greater than a=0.05, the obtained results satisfied the normality test. According to this, for the results of the AUC measure on balanced datasets of the SVM classifier on Group1 and Group2, a t-test can be performed. Two-sample two-tailed t-test assuming unequal variances, i.e., Welch's corrected unpaired two-tailed t-test with a = 0.05 is performed.

```
Welch Two Sample t-test
data:  SVMGroup1$AUC_SMOTE and SVMGroup2$AUC_SMOTE
t = -2.0959, df = 25.304, p-value = 0.04625
```

After performing all the described statistical tests, we obtained a statistically significant difference only in the SVM AUC results for balanced datasets.

The difference is statistically significant on the SVM average accuracy too, but the normality test does not satisfy. All others differences are not statistically significant.

Finally, the results of the second stage of the research show that the cardinality of the datasets has a significant impact only on the performances of the SVM classifier. According to the AUC measure, after balancing training datasets, the linear support vector machine algorithm performs significantly better for Group2 than Group1. This finding raises the question; whether the results would be even better for the linear kernel SVM if the number of observations in the datasets was higher?

According to the data presented in the tables and figures, the results of the second stage of the research show: (i) the impact of the actual number of observations in datasets is most evident in the performances of the SVM classifiers, (ii) the neural network and the RIP classifiers perform relatively more uniformly than SVM and NB, according to the number of instances in a dataset.

6. Conclusions

The research results are showing that in the domain of class imbalanced datasets, data re-sampling has a statistically significant positive influence on the performance of all classifiers, except Naïve Bayes, measured by the AUC measure. At the same time, in the same datasets, the average classification accuracy of all classifiers is statically significantly better when the models are constructed based on original datasets. This is the answer to the first question of interest of this research.

The experimental results suggest that the classifier designer should take into account a trade-off between performance measures. That is, making a classifier better in terms of a particular measure can result in a relatively worse classifier in terms of another. Because of this, the justification of the use of an additional technique for impact reduction of a class imbalance in the classification process depends on the classification goal. Nevertheless, the results demonstrate that if we want to construct a classifier that will be optimized to accuracy as primary performance measure, then we need to use original datasets without balancing.

The impact of the number of dataset instances to classification algorithm performances is most evident on the SVM classifiers after the dataset balancing. This is the answer to the second question of interest of this research. Sensitivity of the SVM can be used as a source for classifier diversity; one of the key factors when designing a multi-classifier system.

We believe that the results of this study can be a guideline for classifier designers and a useful indicator for their researches. An interesting extension to this research would be to explore: (i) the classifier performances, especially of the SVM, on very large datasets and (ii) how dimensionality of imbalanced datasets can have an impact on classifier performances.

References

[1] Akbani, Rehan, Stephen Kwek, and Nathalie Japkowicz. "Applying support vector machines to imbalanced datasets." Machine Learning: ECML 2004. Springer Berlin Heidelberg, 2004. 39-50.

[2] Brennan, P. (2012). A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection, Institute of technology Blanchardstown Dublin, Ireland.

[3] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.

[4] 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.

[5] Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery, 17(2), 225-252.

[6] Dal Pozzolo, A., Caelen, O., Waterschoot, S., & Bontempi, G. (2013). Racing for unbalanced methods selection. In Intelligent Data Engineering and Automated Learning–IDEAL 2013 (pp. 24-31). Springer Berlin Heidelberg.

[7] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 7, 1-30.

[8] Gamberger, D., Lavrac, N., & Dzeroski, S. (2000). Noise detection and elimination in data preprocessing: experiments in medical domains. Applied Artificial Intelligence, 14(2), 205-223.

[9] Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Advances in intelligent computing (pp. 878-887). Springer Berlin Heidelberg.

[10] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on, 21(9), 1263-1284.

[11] Japkowicz, N. (2000). Learning from imbalanced data sets: a comparison of various strategies. In AAAI workshop on learning from imbalanced data sets (Vol. 68).

[12] Japkowicz, N., & Shah, M. (2011). Evaluating Learning Algorithms. Cambridge University Press.

[13] Kotsiantis, S. B., and P. E. Pintelas. "Mixture of expert agents for handling imbalanced data sets." Annals of Mathematics, Computing & Teleinformatics 1.1 (2003): 46-55.

[14] Marusteri, M., & Bacarea, V. (2010). Comparing groups for statistical differences: How to choose the right statistical test?. Biochemia medica, 20(1), 15-32.

[15] Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks, 21(2), 427-436.

[16] Myers, J. L., & Well, A. (2003). Research design and statistical analysis. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc..

[17] Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Systems with Applications, 41(4), 2052-2064.

[18] Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert systems with applications, 39(16), 12605-12617.

[19] Oreški, G., & Oreški, S. (2014, September). An experimental comparison of classification algorithm performances for highly imbalanced datasets. In 25th Central European Conference on Information and Intelligent Systems.

[20] Raeder, T., Forman, G., & Chawla, N. V. (2012). Learning from imbalanced data: evaluation matters. In Data Mining: Foundations and Intelligent Paradigms (pp. 315-331). Springer Berlin Heidelberg.

[21] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[22] Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research, 2, 45-66.

[23] Van Hulse, J., & Khoshgoftaar, T. (2009). Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering, 68(12), 1513-1542.

[24] Wu, G., & Chang, E. Y. (2003, August). Class-boundary alignment for imbalanced dataset learning. In ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC (pp. 49-56).