An experimental comparison of classification algorithm performances for highly imbalanced datasets

Goran Oreški
GO Studio Ltd.
Unska 54, 44324 Jasenovac, Croatia
goreski@gostudio.hr

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

Abstract. *Imbalanced learning data often emerges during the process of the knowledge discovery in data and presents a significant challenge for data mining methods. In this paper we investigate the influence of class imbalanced data on: artificial intelligence methods, i.e. neural networks and support vector machine and on classical classification methods represented by RIPPER and Naïve Bayes classifier. The research is conducted on classification problems and, in purpose of measuring the quality of classification, the accuracy and the area under ROC curve measures are used. For the reduction of the negative influence of imbalanced data, SMOTE oversampling technique is used. All experiments on 30 different data sets, obtained from KEEL (Knowledge Extraction based on Evolutionary Learning) repository, are conducted on original datasets, and repeated on balanced datasets generated using SMOTE technique. The results of the research indicate that imbalanced data have significant negative influence on AUC measure on neural network and support vector machine. The same methods are showing improvement of AUC measure when applied on balanced data, but at the same time, are showing the deterioration of results from aspect of the classification accuracy. RIPPER results are also similar, but the changes are of smaller magnitude, while results of Naïve Bayes classifier show overall deterioration of results on balanced distributions.*

Keywords. imbalanced data, classification learning algorithm, re-sampling technique, 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 with the vast number of cases where the 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 between-class imbalance, i.e. the case when some classes have much more examples than others. By convention, in imbalanced data sets, we call the classes having more examples the majority classes and the ones having fewer examples the minority classes. As well, the class label of the minority class is positive, and the class label of the majority class is negative [7]. The fundamental issue with the imbalanced learning problem is the ability of imbalanced data to significantly compromise the performance of 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 [8].

In recent years there are many scientific papers that address this topic. Most of the papers are focused on finding the best classification algorithm for certain dataset or datasets [3][13], as well as on proposing new techniques for data re-sampling [4][7].

The main goal of the study presented in this paper is to explore the key characteristics of the 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.

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 SMOTE technique. Section 4 describes the experimental design. In Section 5 we provide empirical results with discussion. Section 6 concludes this paper and gives some guidelines for future work.

2 Problem statement and literature review

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

The focus of this research is (1) to analyze the usage justification of additional technique for impact reduction of class imbalance, named SMOTE, in the classification process and (2) to analyze the application impact of this additional preprocessing technique on classification algorithm performances.

The literature in the field of class imbalance is numerous. One of the first studies which brought together the previous research work is the paper Japkowicz [9]. 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; creation of new technique for data balancing [4][7], analysis of the relationship between class imbalance and cost of miss-classification [5], research of different evaluation measures for used models in class imbalance conditions [15], finding the best strategies for establishing the optimal relationship in imbalanced data [6].

According to the topic of this research, in the next section we provide short description of the selected algorithms, whose performances are studied.

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 experiment: 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 short description of SMOTE technique, used for distribution balancing of datasets.

3.1 Neural network

Neural networks (NN) are part of computational and artificial intelligence field and therefore can be classified as artificial intelligence method. There are many different kinds of neural networks and neural network algorithms. The neural network algorithm used in the experiment 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 [14]. This algorithm is a variation of the gradient descent algorithm to find a minimum of an error function in the weight space [11]. 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 [17]. 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 far outnumbers the positive instances, the performance of SVM drops significantly [18].

3.3 RIPPER (Repeated Incremental Pruning to Produce Error Reduction)

As an example of classical algorithmic approach to solving the class imbalance problem, the simple rule induction learning algorithm, RIPPER is used. RIPPER algorithm is a rule induction system which makes use of a divide and conquers 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 [10]. The naive Bayes classifier greatly simplifies learning by assuming that features are independent 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 [16].

3.5 SMOTE

In 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 we 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 re-sampling technique. The 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 Methodological background section. We use the default parameter values in the each case for each algorithm, because our main aim is to highlight the differences between their basic performance, measured with and without SMOTE re-sampling technique, and not to find the best classifier.

4.1 Research procedure description

Initially, 30 different imbalanced datasets are selected from KEEL repository. Each original dataset is presented as the input of four selected learning algorithms. 10-fold cross-validation technique is used in order to create and validate performance of the models. Second procedure, with the SMOTE technique included, was different. In this procedure preprocessing step is added. All datasets are re-sampled, i.e. balanced with SMOTE technique. Balanced datasets are taken as input to 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. All 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.

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 much more majority examples than minority examples. Under this circumstance, accuracy cannot reliably evaluate prediction for the minority class. Thus, more reasonable evaluation metrics are needed. The Area Under the ROC Curve (AUC) is accepted as traditional performance metric in a such situation.

4.2 Statistical comparisons

The research results are verified by statistical tests. The results of each dataset are tested before and after balancing. From statistical point of view, every time, we are comparing the performance of two classifiers on a single domain. Testing was performed by the paired *t* test, one of the most widely used statistical significance measures currently adopted in the context of classifier evaluation. Additional statistical testing was done with nonparametric alternative that is convenient for comparing two classifiers on a single domain; Wilcoxon matched pairs signed ranks test. In order to reduce the likelihood of the type I error, tests were made with the significance a=0.01.

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

5 Results and discussion

The research was conducted on 30 different datasets, obtained from KEEL (Knowledge Extraction based on Evolutionary Learning) repository, with a wide variety of class distributions and with the different number of observations in data sets. In these datasets, the imbalance ratio goes from 9:1 to 41:1, and number of observations goes from 92 to 1829.

In the 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 corresponding *p*-values are

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

Dataset	**NN**		**SVM**		**RIP**		**NB**	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
cleveland-0_vs_4	0,9474	0,9595	0,9536	0,6705	0,8958	0,9711	0,9301	0,9191
ecoli-0-1_vs_2-3-5	0,9754	0,9549	0,9508	0,8852	0,9672	0,9467	0,9098	0,9672
ecoli-0-1_vs_5	0,9792	0,9250	0,9667	0,9375	0,9792	0,9583	0,9792	0,8042
ecoli-0-1-3-7_vs_2-6	0,9929	0,9146	0,9751	0,8221	0,9893	0,9680	0,9502	0,8007
ecoli-0-1-4-7_vs_2-3-5-6	0,9792	0,9137	0,9167	0,9137	0,9762	0,9613	0,9315	0,9137
ecoli-0-1-4-7_vs_5-6	0,9730	0,9669	0,9701	0,9066	0,9458	0,9639	0,9580	0,9337
ecoli-0-3-4-7_vs_5-6	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-0-1-4-6_vs_2	0,9174	0,6976	0,9174	0,3122	0,8974	0,8000	0,4431	0,4146
glass-0-1-5_vs_2	0,9012	0,4767	0,9012	0,1802	0,9302	0,8953	0,4419	0,4070
glass-0-1-6_vs_2	0,9115	0,7708	0,9115	0,2708	0,9427	0,8385	0,4219	0,3906
glass-0-1-6_vs_5	0,9565	0,9728	0,9511	0,8098	0,9946	0,9728	0,9783	0,8641
glass-0-4_vs_5	0,9565	0,9239	0,9022	0,8913	0,9891	0,9891	0,9891	0,4457
glass-0-6_vs_5	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
led7digit-0-2-4-5-6-7-8-9_vs_1	0,9549	0,9300	0,6187	0,8533	0,9617	0,9549	0,8985	0,8262
page-blocks-1-3_vs_4	0,9576	0,9725	0,9661	0,9343	0,9957	0,9873	0,9386	0,9534
shuttle-c0-vs-c4	0,9995	1,0000	1,0000	1,0000	1,0000	1,0000	0,9989	0,9978
yeast-0-2-5-6_vs_3-7-8-9	0,9313	0,8825	0,9084	0,8705	0,9502	0,9133	0,9163	0,9203
yeast-0-2-5-7-9_vs_3-6-8	0,9641	0,9333	0,9691	0,9293	0,9561	0,9622	0,8884	0,7590
yeast-0-3-5-9_vs_7-8	0,9091	0,8399	0,9170	0,7273	0,9289	0,8439	0,5652	0,3439
yeast-0-5-6-7-9_vs_4	0,9375	0,8333	0,9034	0,8182	0,9527	0,8674	0,5473	0,2879
yeast-1_vs_7	0,9346	0,7908	0,9346	0,7691	0,9651	0,9172	0,5163	0,3203
yeast-1-4-5-8_vs_7	0,9567	0,5758	0,9567	0,6335	0,9567	0,8874	0,2063	0,1573
yeast-2_vs_4	0,9689	0,9436	0,9339	0,9339	0,9747	0,9533	0,8677	0,4844
yeast-2_vs_8	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 balanced dataset.

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 NN, RIP and NB classifier 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., differences between paired values are not consistent [12]. Additional statistical test was done with

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

Dataset	NN		SVM		RIP		NB	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
cleveland-0_vs_4	0,7952	0,9428	0,7630	0,6452	0,6611	0,9137	0,8918	0,8856
ecoli-0-1_vs_2-3-5	0,8936	0,9564	0,7500	0,8621	0,9447	0,9333	0,5602	0,9261
ecoli-0-1_vs_5	0,9205	0,9364	0,8000	0,9205	0,9205	0,9318	0,8977	0,8705
ecoli-0-1-3-7_vs_2-6	0,8571	0,9562	0,5000	0,9088	0,8553	0,9836	0,9745	0,8978
ecoli-0-1-4-7_vs_2-3-5-6	0,8949	0,5000	0,5172	0,5000	0,9089	0,8539	0,6347	0,5000
ecoli-0-1-4-7_vs_5-6	0,8935	0,9270	0,8367	0,9128	0,7502	0,9437	0,7567	0,9274
ecoli-0-3-4-7_vs_5-6	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-0-1-4-6_vs_2	0,5000	0,8351	0,5000	0,6250	0,5161	0,891	0,5898	0,6541
glass-0-1-5_vs_2	0,5000	0,7097	0,5000	0,5452	0,6732	0,8896	0,6380	0,5924
glass-0-1-6_vs_2	0,5000	0,7681	0,5000	0,6000	0,7561	0,8318	0,6297	0,6126
glass-0-1-6_vs_5	0,5556	0,9857	0,5000	0,6365	0,9971	0,9857	0,9886	0,9286
glass-0-4_vs_5	0,7778	0,9578	0,5000	0,7416	0,9940	0,9940	0,9940	0,6928
glass-0-6_vs_5	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
led7digit-0-2-4-5-6-7-8-9_vs_1	0,8648	0,9495	0,7796	0,8954	0,8931	0,9262	0,8709	0,8560
page-blocks-1-3_vs_4	0,7098	0,9854	0,7310	0,7476	0,9810	0,9932	0,7666	0,9418
shuttle-c0-vs-c4	0,9959	1,0000	1,0000	1,0000	1,0000	1,0000	0,9994	0,9951
yeast-0-2-5-6_vs_3-7-8-9	0,6965	0,8179	0,5533	0,8067	0,8015	0,8755	0,6747	0,7849
yeast-0-2-5-7-9_vs_3-6-8	0,8722	0,9180	0,8839	0,9113	0,8677	0,9655	0,8841	0,8213
yeast-0-3-5-9_vs_7-8	0,5934	0,7776	0,6067	0,7418	0,6934	0,8243	0,6875	0,6004
yeast-0-5-6-7-9_vs_4	0,7378	0,8377	0,5000	0,7943	0,8162	0,8916	0,7057	0,5708
yeast-1_vs_7	0,5775	0,7796	0,5000	0,7679	0,7488	0,8782	0,7103	0,6364
yeast-1-4-5-8_vs_7	0,5000	0,7146	0,5000	0,6493	0,5000	0,8298	0,5852	0,5596
yeast-2_vs_4	0,8868	0,9600	0,6667	0,9022	0,8900	0,9392	0,8829	0,6877
yeast-2_vs_8	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.*

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 the all 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 that classifiers: NN, SVM and RIP have better average AUC scores on balanced (SMOTE) datasets while NB classifier has better average AUC score on original datasets. Applied statistics, two-tailed paired *t*-test and the two-tailed Wilcoxon's matched-pairs signed rank test, show that the AUC differences within NN, SVM and RIP classifiers are statistically significant before and after the datasets balancing. Only NB classifier has better average AUC score on original datasets than “SMOTED” but this difference is not statistically significant.

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

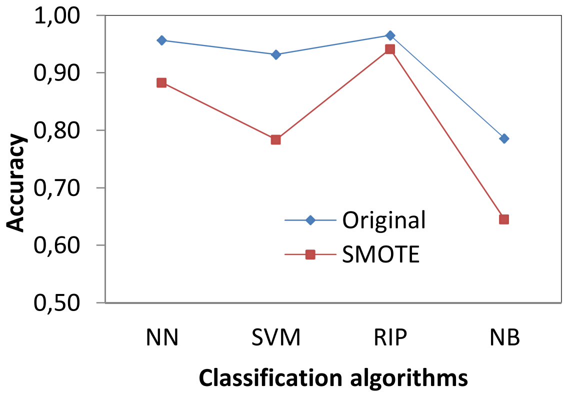

Figure 1. Comparison of the average accuracy of classifiers on original and balanced datasets

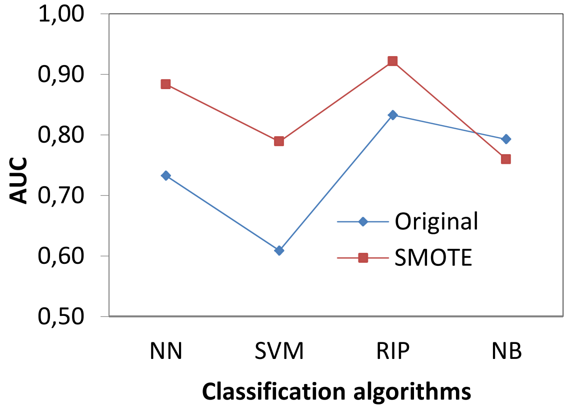

Figure 2. Comparison of the average AUC of classifiers on original and balanced datasets

The results from this empirical study indicate that Ripper classifier is able to cope comparatively well with pronounced class imbalances. At this classifier, balancing of the sets has 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 NN classifier.

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 AUC measure, without the balancing the linear support vector machine algorithm performs the poorest. This finding is consistent with 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 of the research are showing that imbalanced data have significant negative influence on AUC measure at the neural network classifier and, even more, at the linear support vector machine. The same methods are showing improvement of AUC measure when applied on balanced data, but at the same time, are showing the deterioration of results from aspect of classification accuracy. The performances of Ripper classifier are positively correlated with NN and SVM, but the changes are of smaller magnitude, while results of Naïve Bayes classifier show overall deterioration of results on balanced distributions.

6 Conclusions

The research results are showing that in domain of class imbalanced datasets, re-sampling SMOTE technique has statistically significant positive influence on performance of all classifiers, except Naïve Bayes, measured by AUC measure. In the same time, on 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 on the second question of interest of this research.

Unfortunately, because of the inductive nature of the problem, first question of interest of this research is not fully answered. Instead, the classifier designer should take into account results of this study and 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 using the additional technique for impact reduction of a class imbalance, named SMOTE, in the classification process depends of the classification goal.

We believe that results of this study can be the guideline for classifier designers and a useful indicator for future research. An interesting extension to this research would be to explore the effect of the actual number of observations in datasets to performances of classifiers.

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] 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.

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

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

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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).

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

[18] 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).