Cost-Sensitive Learning from Imbalanced Datasets for Retail Credit Risk Assessment

Stjepan Oreški [1], Goran Oreški [2]

[1] *Bank of Karlovac, I.G.Kovacica 1, 47000 Karlovac, Croatia*
[2] *GO Studio Ltd, Unska 54, 44324 Uštica, Croatia*

***Abstract* – In the present study we propose a new classification technique based on genetic algorithm and neural network, optimized for the cost-sensitive measure and applied to retail credit risk assessment. The relative cost of misclassification, which properly accounts for different misclassification costs of minority and majority classes, is used as the primary evaluation measure. The test of the new algorithm is performed on Croatian and German retail credit datasets for seven different cost ratios. An empirical comparison with others in the literature presented models demonstrates the potential of the new technique in terms of misclassification costs.**

***Keywords* – genetic algorithm; neural network; credit risk assessment; imbalanced datasets; misclassification cost.**

1. Introduction

Imbalanced datasets are common in many real-world domains, including finance, medicine, telecommunications, ecology, and biology. Such datasets can be considered as one of the major issues of the data mining process. A dataset is imbalanced if the classification categories are not approximately equally represented [1], [2]. It should be noted that the minority class is usually of the highest interest because, when misclassified, it represents the largest cost [3]. Accordingly, class imbalanced datasets are very often related to cost-sensitive learning. Therefore, accurate classification of minority class instances is very important.

Most of the published literature indicates that standard classification algorithms on imbalanced datasets suffer a significant loss of performance. It is inherent to the mentioned algorithms to often be biased in favor of the majority class, otherwise known as the "negative" class. There is a higher classification error rate for minority class instances, known as the "positive" class. Over the past years, many techniques have been proposed to address this type of classification problem, either through: (i) the data, (ii) the algorithmic, or (iii) the hybrid (combined) approach. All these techniques address, in their own way, the problem of class imbalance and contribute to improving the performance of standard classifiers when they are applied to the class-imbalanced data.

Classifiers are generally based on machine learning algorithms, and their performance is usually estimated according to the predictive accuracy of constructed models. However, this measure is not appropriate when the distributions are imbalanced and/or the costs of different errors vary greatly [4]. In support of this thesis, Chawla et al.[5] state that a typical mammogram dataset can contain 98% normal pixels and 2% abnormal pixels. The simple strategy of guessing, which always chooses the majority class, would provide a predictive accuracy of 98%. However, to achieve the goal of classification, the nature of application requires a fairly high rate of predictive accuracy of the minority class and allows for a small error rate in the majority class. Therefore, as is obvious, the standard measure of predictive accuracy is not appropriate in such circumstances. Many other measures showing the performance of the classifier are proposed in the literature. Their selection should be in accordance with the purpose and objectives of classification.

In this introduction, we have mentioned some of the issues related to the classification in class and cost imbalanced datasets. One of the domains that is

DOI: 10.18421/TEM71-08
https://dx.doi.org/10.18421/TEM71-08

Corresponding author: Stjepan Oreški,
Bank of Karlovac, Karlovac, Croatia
Email: stjepan.oreski@kaba.hr

Received: 14 November 2017.
Accepted: 24 January 2018.
Published: 23 February 2018.

sensitive to this type of issues is banking, especially credit risk assessment, because in many banks models of artificial intelligence take the role of decision-makers in the process of loan approval. Thereby, a poor credit risk assessment model, which does not take into account differential error misclassification costs, could lead to sub-optimal capital allocation [6]. In order to construct cost-sensitive models, the objective of this study is exploring the impact of resampling techniques in combination with the feature selection technique to the classification results measured by the relative cost of misclassification. As a classifier, on which the impact of resampling techniques will be explored, the hybrid genetic algorithm with neural networks (HGA-NN), which performs feature selection and classification simultaneously, will be used [7]. Because of the pronounced asymmetry of the misclassification costs and their contribution to the financial performance of the financial institution, in the analysis of the classifier performance, the emphasis is on the analysis of the misclassification cost. With respect to the analysis, the algorithm optimized for the relative cost of misclassification will be proposed.

The study is organized as follows. Section 2 describes, in more detail, the problem of class imbalance in credit risk assessment and reviews the literature related to this problem. Section 3 describes techniques of class imbalance problem solving. Special emphasis is given to specifics of evaluation measures and techniques of validation to be used in terms of class imbalance. Section 4 describes the new technique for attributes selection and classification. Section 5 presents the results of the experiments, the performance evaluation and their comparison with the results presented in the literature. Section 6 presents this study’s conclusions.

2. The Problem Description and the Literature Review

Data with class imbalance often occurs in the field of classification. The main characteristic of this classification problem type is that one class’s examples significantly surpass the number of instances of other classes [3], [8]. The imbalance is expressed through the imbalance ratio (IR), which is defined as the ratio of the number of cases in the majority class to the number of examples in the minority class. In most cases, the class imbalance problem is associated with a binary classification, but it also appears in multiclass problems, in which there may be more minority classes, causing an even greater problem in the classification [3].

It is important to note that it is usually more difficult to obtain instances of real data for the minority class and that the collection of such data is associated with significant costs. In credit risk assessment, the minority class represents bad loans. The costs of misclassifying a client that subsequently fails (bad loan) are very different to the costs of misclassifying a client that does not fail. In the first case, the lender can lose up to 100% of the loan amount while, in the latter case, the loss is just the opportunity cost of not lending to that client. Therefore, the loan approval process aims to reject bad applicants and instances of bad loans are less frequently in real data.

Because most of the standard algorithms for machine learning assume a balanced set of training data with approximately uniform cost of incorrect classification, in imbalanced sets, these algorithms generate suboptimal classification models. Usually, the majority cases are well classified, while minority instances are more often incorrectly classified. Therefore, the algorithms that achieve the best results with a balanced set do not necessarily achieve the best performance with imbalanced classification sets [9].

Series of scientific studies with the aim to solve such problems is presented in the literature. Chawla et al. [5] state that their technique SMOTE can achieve better classifier performance (in ROC space) than it is achieved by changing the loss ratio with RIPPER techniques or changing the class priority with a naive Bayes classifier. Dal Pozzolo et al. [10] perform parallel testing of alternative strategies on a subset of the data and progressively abandon alternatives that are significantly worse. They conclude that the best strategy depends on the applied algorithm and the dataset used. Van Hulse and Khoshgoftaar [11] infer that a simple resampling technique, such as random undersampling, is generally the most effective to improve models’ performance. Their experiment is designed on the basis of imbalanced and noisy data. Chawla et al. [12] analyze the impact of relationships between classes on the misclassification cost. To optimize the relationships between classes before model construction they propose the envelope technique, which finds the necessary (optimal) relation between classes by repeating the resampling and optimization of evaluation functions such as F-measure, AUC, cost, cost-curves and cost dependent F-measure.

Datasets that have a highly imbalanced class distribution represent a fundamental challenge in machine learning, not only in terms of construction of the model but also in terms of ways to measure the quality of the constructed models. There are many different measures of model evaluation used in class imbalance conditions, each with its own bias. There are also different strategies of cross-validation. Selected evaluation measures and cross-validation strategy must be consistent with the problem to be analyzed [13], and their characteristics should be well known. Although there are different technique proposals to solve the problems of class imbalance, it can be observed from the literature review that a good technique has to consider: (1) collected data, (2) the classification method and (3) the performance measure.

Because of the great diversity of the presented techniques and algorithms, as well as the diversity of the application domains, it is difficult to simply classify all existing approaches for reducing problems related to cost-sensitive classification on imbalanced data. From the literature review [14] [15], [16], it can be concluded that there is a relatively great interest in the study of this problem, but in the area of the credit risk assessment this subject is not adequately researched. The need exists, because instances of bad loans are less frequently in real data and the costs of misclassifying bad and good clients are very different. As well as in the previous research efforts, we did not find studies that research feature selection and resampling techniques in combination with misclassification costs. Therefore, the research presented below seeks to cover a perceived gap by exploring the impact of resampling techniques, combined with the feature selection technique, on the classification results, primarily based on the misclassification cost. The study creates a new technique optimized for the relative cost of misclassification in the domain of retail credit risk assessment. Thereby, small differences in model power can lead to significant economic impact for the users. Hence, research in this area assumes greater significance because a poor credit risk model could lead to sub-optimal capital allocation.

3. Class Imbalance Problem Solving Techniques

Data mining and machine learning consider class imbalance mainly in two ways. One is by assigning different weights to the training examples (algorithmic approach), and the other is by resampling the original data (data approach) [5]. These approaches can also be used in combination with each other, i.e., the hybrid approach. Conclusions concerning which approach is better are not always the same, but, in most cases, the efficiency of data approaches was better [5], [11]. The data approach is more present because the technique of this approach is independent of the used classifier and can be easily applied to any problem. The vast majority of authors use resampling techniques to mitigate the problem of class imbalance [17] by balancing the number of samples in the minority and majority classes [2]. Regularly, resampling is carried out as long as the classes are not approximately equally represented.

Random oversampling (ROS) of the minority class is the simple strategy of using random replication of positive examples to balance the class distribution. This technique increases the overfitting likelihood of minority class examples because it makes exact copies of the minority examples. To overcome the overfitting of minority class examples, Chawla et al. [5] have proposed a managed technique named Synthetic Minority Oversampling Technique (SMOTE). This technique generates artificial positive examples by interpolating attribute values of the existing closest examples. It does this by finding the *k* nearest neighbors of a minority class and then generating new synthetic examples in the direction of some or all of the nearest neighbors, which depends on the amount of the requested new examples [5]. Random undersampling (RUS) of the majority class is the resampling technique that removes the majority class examples from the sample as long as a minority class does not reach a defined percentage of the majority class.

RUS and ROS have different drawbacks. RUS can potentially remove some important examples of the majority class, and ROS can lead to the overfitting of minority class examples [1]. López et al. [3] also concluded that the more sophisticated techniques are less general, were developed specifically for a particular set of problems, and, when compared to a large number of reference problems, might provide inferior performance.

Sampling techniques are frequently used techniques for solving problems related to class imbalance and the different costs of incorrect classification, but they are not the only ones used. Techniques that fall into the algorithmic approach are also used to solve this type of classification problem [3]. The central issues in the algorithmic approach for cost-sensitive learning are: (i) how to take into account the costs of incorrect classifications during model construction and (ii) how to adapt the model to user requirements in the operation phase. The algorithmic approach affects the classification results in some of the following ways, i.e., through; (1) changes in the underlying algorithm, (2) unequal weight given to the class instances to guide a classifier to pay much more attention to the minority class, or (3) changes in the classification threshold.

Hybrid approaches include techniques on the data and algorithmic level, as well as their combinations, so that the created model minimizes the misclassification costs. With the inclusion of hybrid techniques, the construction process of a classification model becomes more complex. Overall, the idea of the cost-sensitive classification is not to make the least amount of errors in the classification or to achieve maximum accuracy, but to construct a model that will produce the least possible misclassification cost. Objectives, such as error minimization and cost minimization, are not necessarily the same; moreover, in imbalanced sets, they are usually not.

3.1. *Evaluation measures*

The quality of the constructed classification model depends on the quality of the classification algorithm, the data quality and the selected evaluation measures. Specifically, we will not obtain the same model quality picture when measuring this quality according to measure A or B. Although each measure has shortcomings, using an array of measures provides a richer picture of model quality than if we used only a single measure. Which of the result evaluation measures will be used as the main one in a particular case depends on the nature of the problem and the objectives of classification.

The results of classification and validation can be shown in a confusion matrix (CM), which is a useful tool for analyzing how well a classifier can recognize tuples of different classes [18]. A confusion matrix for two classes is shown in Table 1.

Table 1. Confusion matrix for a two-class problem

		Predicted result		Recognition rate
		Default	Non-default	
Real	Default	true positives (TP)	false negatives (FN) (Type I error)	Sensitivity (Recall)
	Non-default	false positives (FP) (Type II error)	true negatives (TN)	Specificity
		Precision		Accuracy (%)

In order to compare two different models, we can compute diverse measures from the confusion matrix. Classification accuracy is the most common measure of performance that directs machine learning algorithms and, according to the confusion matrix (Table 1), is defined as:

$$Accuracy\ (\%) = (TP + TN) / (TP + FP + TN + FN) * 100. \tag{1}$$

Predictive accuracy might not be appropriate when the data are imbalanced and/or the costs of different errors vary markedly [1]. When the specified conditions are met, the Area Under the ROC Curve (AUC) can be used as a measure of model quality. It combines the FPR and the TPR ratios into one single measure. The ratios TPR (or sensitivity, or recall) and FPR (or 1-specificity) are calculated according to the following equations:

$$TPR = TP / (TP + FN), \tag{2}$$

$$FPR = FP / (FP + TN). \tag{3}$$

The ROC curve best shows the relationship between the TPR and the FPR ratios. The X-axis represents the FPR and the Y-axis represents the TPR [19]. The ROC curve treats the costs of a type I error (classifying a subsequently failing client as non-failed) and a type II error (classifying a subsequently non-failed client as failed) the same [6]. Accordingly, the AUC is the classifier quality measure independent of the selected decision-making criterion and a priori probability [1]. The independence of the a priori probability and decision-making criterion is not always desirable. Indeed, if we want to construct a classifier that will be best adapted to the specific decision-making criterion, then we need the measure that would best express that criterion. Based on that measure, we will make the right decision about the best classifier. For a certain class distribution and costs, the classifier with the best AUC can be suboptimal. Therefore, in addition to the AUC, we use the F-measure and the relative cost of misclassification as additional measures of the classifier quality.

While the ROC curve plots the relationship between the TPR and the FPR ratios, the F-measure is the ratio between the values TP, FP and FN. The F-measure balances the relationship between precision and recall, and the result is a number that reflects the "goodness" of the classifier for the minority examples. The F-measure is the harmonic mean of precision and recall and tends towards the lower of the two [20]:

$$F\text{-}measure = (2 * precision * recall) / (precision + recall), \tag{4}$$

where the equation for recall is equivalent to the TPR and the equation for the precision calculation is

$$precision = TP/(TP+FP). \tag{5}$$

There is a weighted measure of precision and recall:

$$F\text{-}_{\beta} = ((1+ \beta^{2}) * precision * recall) / (\beta^{2} * precision + recall), \tag{6}$$

where β corresponds to the relative importance of recall versus precision.

In this study, the emphasis is on the misclassification cost. The total relative cost of misclassification (RC) will be calculated according to the following equation [21] :

$$RC = \alpha\ (P_{I}C_{I}) + (1\text{-}\ \alpha)\ (P_{II}C_{II}), \tag{7}$$

where α is the probability of being a ‘bad’ client, P_I is the probability of a type I error, C_I is the relative cost of the type I error, P_{II} is the probability of the type II error, and C_{II} is the relative cost of the type II error. The RC of each model is computed for seven cost ratios, while the best model for each ratio is the model with the lowest RC value. While the accuracy, as a conventional performance measure, ignores the inherent costs of type I and type II errors to profits, RC, on the other hand, takes into account costs of type I and type II errors, and provides a more suitable risk-based performance measure. Accordingly, RC takes better into account the objectives of the lending company.

3.2. Validation techniques

As we noted, in imbalanced datasets, it is particularly important to use appropriate measures for the quality of the constructed model. Almost equally important is to select the most appropriate technique for model performance validation. Classifier performance validation is a necessary procedure to assess how a classifier will perform when it classifies new instances. The way this task is performed has a direct influence on the analysis of the quality of the constructed model [22].

Applying a random division of the instances over training and test folds may result in a problem known as dataset shift. The problem of dataset shift is defined as the case where training and test data follow different distributions. This is a common issue that can affect all types of classification problems. The issue is especially relevant when dealing with imbalanced classification because a misclassified example of the positive class can create a significant drop in performance. Stratified *k*-fold cross-validation is a technique used usually for assessing how a classifier will perform when classifying new instances of the task at hand. This avoids prior probability shift because, with an equal class-wise distribution on each fold, the training and test set will have the same class distribution.

After the application of a resampling technique for balancing the imbalanced dataset, the sample is not representative in relation to the population. The shift is intentionally induced. Therefore, the best representative of the entire data population is the original sample, before balancing. Regarding this fact, a stratified 10-fold cross-validation technique is used for learning on the balanced dataset, and the original dataset is used for the final model performance validation.

4. Model Development

The class imbalance and different misclassification costs represent significant challenges for classification and have significant impact on some of the performance measures. From the results presented in [7], it is apparent that the HGA-NN algorithm achieves excellent results in terms of classification accuracy in retail credit risk assessment. To achieve equally good results according to the relative costs of misclassification, new techniques will be incorporated in this algorithm. The newly created algorithm optimizes performances in terms of the average relative cost of misclassification, as well as in relation to other measures of the classification quality inherent to class imbalance.

4.1. The extended HGA-NN algorithm

The HGA-NN technique, presented in [7] utilizes the earlier experience of experts and the efficiency of fast algorithms for feature ranking, as well as the optimization capabilities of a GA. This three-step hybrid algorithm includes search space reduction, refining of the reduced feature subset, and incremental stages. Search space reduction quickly removes most of the irrelevant features. Refining of the reduced feature subset then further examines the reduced feature set. An incremental stage additionally improves the model’s performance.

After a careful evaluation a lot of experimental results, some practical implementations and checking the findings in the literature, we designed the technique optimized for the cost-sensitive measure. Figure 1 shows the extended HGA-NN technique which optimizes the performance in relation to the classification quality measures inherent to class and cost imbalance.

Our preliminary experiments and Marqués et al. [23] indicate that search space reduction should be applied on the original dataset and that only after search space reduction can we use some of the resampling techniques. Accordingly, resampling with some of the previously described resampling techniques, such as ROS, RUS or SMOTE, was performed after search space reduction and before the reduced feature subset refinement. Thus, the classification algorithm, which calculates the goodness of individuals in a population, uses a balanced dataset. This provides an environment that minimizes the weaknesses of most algorithms for classification, i.e., their bias to the majority class, while providing favorable conditions for fast feature selection algorithms.

```
Input:   originalDataset
         expertFeatureSubset     // if exists
Output:  modelPerformance

// Search space reduction
reducedFeatureSet.add(expertFeatureSubset)
featureSelector = {GA-NN, InformationGain, GainRatio, Gini, Correlation,
                  ForwardSelection}

// GA-NN based feature selection and parameter optimization
bestPerformance = performGANN(originalDataset)
reducedFeatureSet.add(bestPerformance.featureSubset())
bestParameters = bestPerformance.getParameters()
featureMax = bestPerformance.featureSubset().size

// Feature selection for all featureSelector except GA-NN
for (int i = 1; i < featureSelector.length; i++) {
     reducedFeatureSet.add(featureSelector[i].select(originalDataset,
                      featureMax))}
samplingTechnique = {ROS, RUS, SMOTE}
for (int i = 0; i < samplingTechnique.length; i++) {
     // Resampling
     balancedDataset = reSampling(samplingTechnique[i], originalDataset,
                          reducedFeatureSet)
     // Reduced feature subset refinement
     do {
           population = createInitialPopulation(balancedDataset,
                  reducedFeatureSet)
           performances = performanceCalculation(balancedDataset,
                    population, bestParameters)
           bestPerformance = performances.getBest()
           while (NOT convergenceCriterion) {
           population = generateNewGeneration(population, performances)
           performances= performanceCalculation(balancedDataset,
                    population, bestParameters)
           if (bestPerformance.compare(performances.getBest())) {
                  bestPerformance = performances.getBest()
           }
           } // end while
           if (generationsWithoutImproval <= 2) {
           // Add designated solution to initial solutions
              reducedFeatureSet.add(bestPerformance.featureSubset())
           }
     // Incremental stage control
     } while (generationsWithoutImproval <= 2)
  } // end for
  // Final model validation
  filteredOriginalData = filter(originalDataset,
                   bestPerformance.featureSubset())
  modelPerformance = modelValidation(filteredOriginalData,
                 bestPerformance.getModel())
  return modelPerformance
```

Figure 1. Pseudo-code of the extended HGA-NN technique

In doing so, regardless of the applied resampling technique, the resampling produces class balance in the learning data. The new algorithm, in all cases, first obtains a balance of 50:50 in the learning data, which is theoretically the best ratio for most algorithms for classification. Class balance is established because the matrix of the costs is unknown. Establishing a relationship other than 50:50, without in an advance defined misclassification costs, with a conceptual aspect is not justified. Defining the misclassification costs is a very challenging and important task in real world financial environments and is outside the scope of this paper. With the matrix of costs, the banks determine their own attitude to risk. Ultimately, cost optimization and capital allocation depend significantly on these values. By applying the described resampling techniques, the initial objective, i.e., the equal treatment of different classes in the model construction, is reached. To assess how the model will perform when it classifies new instances, final model performance validation is performed on the representative sample for the population, i.e., on the original dataset. In addition, after the model construction, we can use the threshold-moving technique to adjust the output threshold toward inexpensive classes such that positive (high-cost) samples are unlikely to be misclassified. With this approach, we can relatively quickly additionally optimize model results in accordance with a specific (a posteriori defined) matrix of costs to obtain the best results for the target cost ratio.

Table 2. Summary of parameters for the HGA-NN and extended HGA-NN used for a Croatian and German dataset

Parameter	Set up for dataset	
	Croatian	German
Population initialization		
population size	50	
initial probability for a feature to be switched on	0.6	
maximum number of features	12	16
minimum number of features	5	6
Reproduction		
Fitness measure	accuracy	
Fitness function	neural network	
the type of neural network	multilayer feed-forward network	
network algorithm	back-propagation	
activation function	sigmoid	
the number of hidden layers	1	
the size of the hidden layer	(number of features + number of classes) / 2 +1	
training cycles	500	50
learning rate	0.6	0.29
momentum	0.2	0.43
selection scheme	tournament	
tournament size	0.05	
dynamic selection pressure	Yes	
keep best individual	Yes	
mutation probability	0.1	
crossover probability	0.9	
crossover type	one point	uniform
Condition for completion		
maximal fitness	Infinity	
maximum number of generations	10	
use early stopping	No	

The algorithm shown in Figure 1 is constructed using the Rapid Miner 5.1.15 and Weka 3.6.10 tools with the parameters shown in Table 2. Parameters are shown separately for a Croatian and German dataset. Rows in the table are merged for parameters that have the same values for both data sets. The parameters are not changed throughout the experiment for the HGA-NN and extended HGA-NN techniques. Extensions of the Rapid Miner tool were necessary to execute the algorithm. The standard Rapid Miner genetic algorithm has been extended so that it can accept, as part of its own initial population, the initial solutions generated by the other techniques and the domain's experts. This extension has also enabled the introduction of the incremental stage of the algorithm.

By keeping the same parameters for the HGA-NN and extended HGA-NN, the differences between the experimental results will be affected by the additional techniques that mitigate the impact of class imbalance and different misclassification costs on classification results; thus, their contribution to the results will be easy to quantify.

4.2. *Results evaluation and comparison*

Comparison of the results obtained by using the standard HGA-NN technique and the results obtained by using the extended HGA-NN technique is possible because the parameters are not changed throughout the experiments. However, additional techniques that mitigate the impact of class imbalance and different misclassification costs on classification results are added. The above shows that the differences in the results are exclusively the result of the technique's extension. To determine the overall quality of the new technique presented here, a comparison of its results and the results of other techniques reported in the literature will be performed. It is justified to expect that banks will want to optimize the cost function for some ratio and not the accuracy function. Because the best ratio for each bank is determined by the banks themselves, the most effective model for each bank depends on this. Because this is unknown, a comparison procedure for seven cost ratios will be presented to select the best model.

Two statistical tests are considered to be most suitable for testing the existence of statistically significant differences in the results of several classifiers for several independent samples. The first is a very well-known parametric test: the analysis of variance (ANOVA) for repeated measures. The second is a rarely used, non-parametric alternative to the above test: the Friedman test [24]. Demšar [25] states that ANOVA can be conceptually inadequate and statistically uncertain because, as a parametric test, it is based on different assumptions (normality, homogeneity of variance) that often are not met due to the nature of the problem. The Friedman test, as a non-parametric version of ANOVA for repeated measures, does not have these limitations [25]. The price for this freedom is paid by a lower power Friedman's test compared with the parametric ANOVA. The Friedman test statistics are based on the average performance ranks (R) of classification algorithms for all datasets and are calculated as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right],$$
$$\text{where} \quad R_j = \frac{1}{N}\sum_i r_i^j. \qquad (8)$$

In Eq. (8), N is the number of different datasets (the ratio of costs) for which measurements were conducted, k is the total number of classification algorithms, and r_i^j is the ranking of the j-th algorithm of the i-th dataset. Statistics for the Friedman test are a chi-square with [k-1] degrees of freedom [25].

Hypotheses with the Friedman test are:

$\mathrm{H_0}$: There is no difference in the distributions of ranks in repeated measurements.

$$\mathrm{H_0}: \mathrm{M_1} = \mathrm{M_2} = \cdots = \mathrm{M_k}$$

$\mathrm{H_A}$: The distributions of ranks in repeated measurements are different, i.e., at least one equality is not satisfied.

If Friedman's test gives a significant *p*-value, we perform the Nemenyi post hoc test. According to the Nemenyi post hoc test, the performance of two classifiers is significantly different if their average ranks differ by at least a critical difference (CD):

$$CD = q_{\alpha,\infty,k}\sqrt{\frac{k(k+1)}{12N}}, \qquad (9)$$

where *N* is the number of different datasets, k is the total number of classification algorithms, and the value $q_{\alpha,\infty,k}$ is based on the critical values used for the Turkey test [25].

5. Empirical Analysis

In this section, the objective is to analyze the classification results of the new technique in the domain of retail credit risk assessment. Classification performances are measured by various measures of performance, focusing on the relative misclassification costs. Thereby, small differences in model power can lead to significant economic impact for the lending institution. The analysis was performed only on two datasets, i.e., Croatian and German, because lending companies don't present publicly their own retail datasets. As a result of such circumstances, a cost-sensitive classification on imbalanced data in the domain of the retail credit risk assessment is not adequately researched.

5.1. Description of the experimental datasets

Two real-world credit data sets have been taken to explore the quality of the new technique in the classification of imbalanced datasets in retail credit risk assessment. The total number of randomly selected instances in the Croatian dataset is 1000, including the 750 who successfully fulfilled their credit obligations, i.e., good credit customers, and 250 who were late in performing their obligations and therefore are placed in a group of bad credit customers. The imbalance ratio is 3:1. The total number of features is 35, including 33 regular features and 2 (id, label) special features. The regular features contain 21 integers and 12 real values. The class label feature is binominal. The Croatian dataset have been described in detail in [21].

The German credit dataset comprises 700 instances of creditworthy applicants and 300 instances of bad credit applicants. The imbalance ratio is 7:3. It contains 30 regular features of the integer data type and 2 (id, label) special features and can be viewed at <http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/download-course-materials/>. All of the features with descriptive statistics are shown in [7].

5.2. Results on Croatian dataset

The results in Table 3 show that, when the HGA-NN technique is used, the obtained model accuracy is higher; however, other performance measures show worse results. When the extended HGA-NN techniques are used, we can see that the HGA-NN ROS provides the best results for all other performance measures, except for accuracy.

Table 3. Results of techniques on the Croatian dataset

Technique	Accuracy	AUC	F-score	F-β	TP	FN	FP	TN
HGA-NN	**82.9**[a]	0.7447	0.6274	0.5955	144	106	65	685
HGA-NN RUS	76.1	0.7767	0.6283	0.7251	202	48	191	559
HGA-NN ROS	78.7	**0.7953**[a]	**0.6559**[a]	**0.7414**[a]	203	47	166	584
HGA-NN SMOTE	76.0	0.7827	0.6330	0.7372	207	43	197	553

[a]The best result for each measure.

As previously stated the misclassification cost of the minority class is usually higher than the cost of incorrect classification of the majority class. Therefore, the highest classification accuracy is not a cost-optimal result. Optimal cost depends on accuracy but also on the relation between the cost of type I and type II errors, which depends of economic cycles, capital availability on the market, bank preferences and other circumstances. In terms of cost, the prediction accuracy deterioration is justified as long as the reduction in cost, due to the reduction of type I errors, results in a lower increase in cost due to an increase in type II errors. The misclassification cost may be optimal with lower accuracy.

Since the costs of type I and type II errors are context specific, it is difficult to evaluate credit risk models based on one ratio alone. According to Eq. (7), the relative cost of misclassification is calculated and presented in Table 4 for each model and for seven ratios of type I and type II errors. For each ratio, the best model is the one with the lowest relative cost value. Table 4 also shows that the most accurate classification is the best classification for the bank, with respect to the relative cost of misclassification, only when the cost ratio of type I and II errors is equal. Already in the cost ratio 2:1 of type I and II errors, the most accurate classification may not be the most cost favorable. As the cost ratio increases, the basic, most accurate technique becomes worse, in terms of cost, in comparison with new techniques that have used some extensions.

Table 4. Relative misclassification costs (RC) comparison on the Croatian dataset

Technique	Cost ratio (C_I:C_{II})						
	1:1	2:1	3:1	4:1	5:1	8:1	10:1
HGA-NN	**0.1710**[a]	0.2770	0.3830	0.4890	0.5950	0.9130	1.1250
HGA-NN RUS	0.2390	0.2870	0.3350	0.3830	0.4310	0.5750	0.6710
HGA-NN ROS	0.2130	**0.2600**[a]	**0.3070**[a]	**0.3540**[a]	**0.4010**[a]	0.5420	0.6360
HGA-NN SMOTE	0.2400	0.2830	0.3260	0.3690	0.4120	**0.5410**[a]	**0.6270**[a]

[a]The best result for each cost ratio.

If the bank wants to optimize its loan approval process with a cost ratio other than 1:1, the results of this experiment indicate that they should then use some type of cost-sensitive learning. In our case, the Croatian dataset results show that extended HGA-NN techniques give, from a cost perspective, better results than the standard technique. In addition, the HGA-NN ROS technique shows the best results for cost ratios of 2:1, 3:1, 4:1 and 5:1, and the HGA-NN SMOTE technique shows the best results for cost ratios of 8:1 and 10:1. Generally, cost-sensitive learning results in less average misclassification cost than the traditional, cost-insensitive approach. Since our focus here is on the impact of model's performance on bank profitability, the best model for each ratio is the model with the lowest RC value.

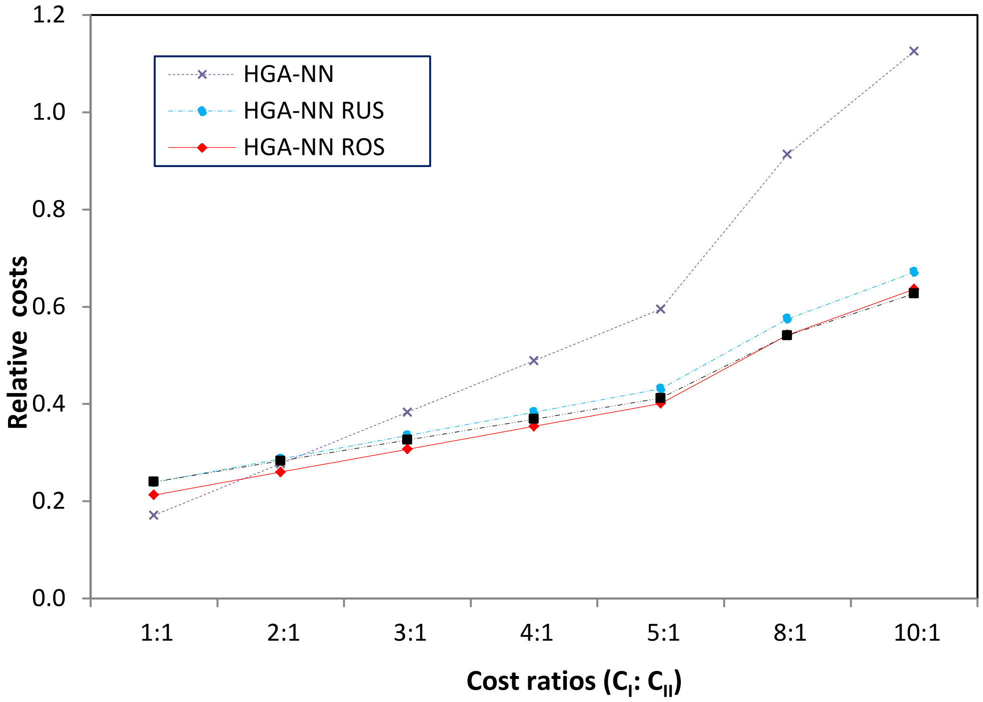

Figure 2. Comparison of relative costs of misclassification for the Croatian dataset

Figure 2 clearly depicts that the standard HGA-NN technique has achieved the best accuracy of prediction, evident from the lowest relative cost for the cost ratio of 1:1. The most accurate model becomes marginally good, from a cost perspective, for the cost ratio of 2:1, and for all cost ratios above 2:1, the model that is constructed without the use of resampling techniques in the Croatian dataset gives poorer results, i.e., higher relative costs. The cost line for this model is the steepest, which means that, with an increase in the relative cost ratio, the model shows the fastest growth in total relative costs.

5.3. Results on German dataset

From the results shown in Tables 5 and 6, it can be observed, as in the Croatian dataset, that the model with the highest accuracy is obtained with the HGA-NN technique; however, for this model, lower (worse) values of other performance measures are reported. Looking at the individual extended techniques, we can see that HGA-NN ROS gives the best results for all other performance measures, except for accuracy.

Table 5. Results of techniques on the German dataset

Technique	Accuracy	AUC	F-vrij.	F-β	TP	FN	FP	TN
HGA-NN	**78.6**[a]	0.7148	0.6008	0.5606	161	139	75	625
HGA-NN RUS	69.5	0.7317	0.6183	0.7269	247	53	252	448
HGA-NN ROS	69.6	**0.7495**[a]	**0.6355**[a]	**0.7641**[a]	265	35	269	431
HGA-NN SMOTE	69.6	0.7419	0.6284	0.7480	257	43	261	439

[a] The best result for each measure.

Table 6. Relative misclassification costs (RC) comparison on the German dataset

Technique	Cost ratio (C_I:C_{II})						
	1:1	2:1	3:1	4:1	5:1	8:1	10:1
HGA-NN	**0.2140**[a]	0.3530	0.4919	0.6309	0.7699	1.1869	1.4649
HGA-NN RUS	0.3050	0.3580	0.4110	0.4640	0.5171	0.6761	0.7821
HGA-NN ROS	0.3040	**0.3390**[a]	**0.3740**[a]	**0.4091**[a]	**0.4441**[a]	**0.5491**[a]	**0.6191**[a]
HGA-NN SMOTE	0.3040	0.3470	0.3900	0.4330	0.4760	0.6050	0.6909

[a]The best result for each cost ratio.

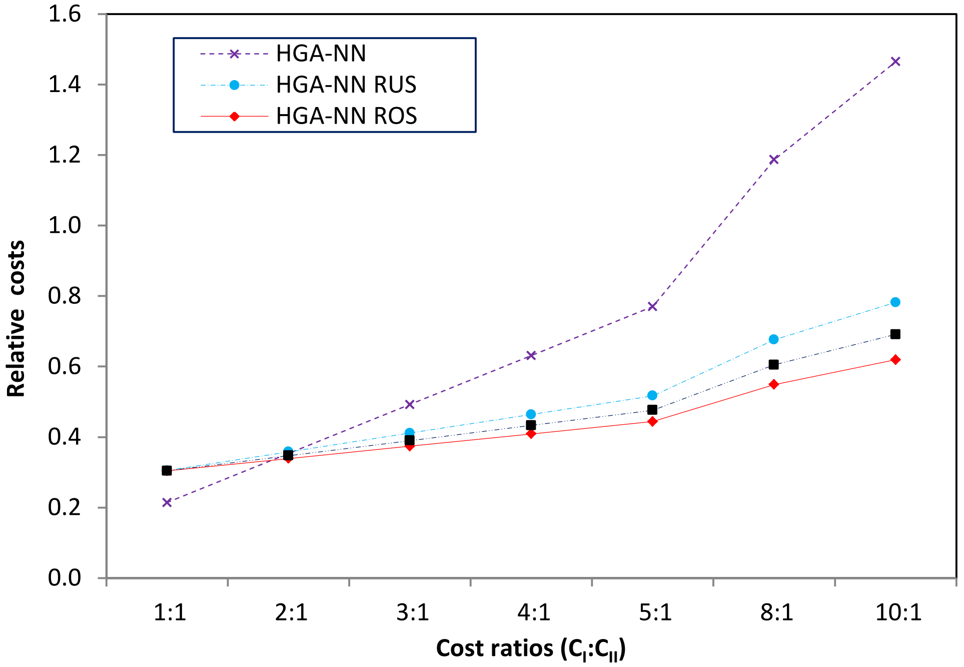

Figure 3. Comparison of relative costs of misclassification for the German dataset

Figure 3 shows the comparison of the relative misclassification cost of the HGA-NN technique and the extended HGA-NN techniques for all constructed models of the German dataset. From the diagram, it is clear that the HGA-NN achieves the highest prediction accuracy. However, it is evident that this technique achieves the lowest relative cost only for the cost ratio of 1:1. The diagram also shows that the most accurate model becomes marginally good, from a cost perspective, with the cost ratio of 2:1, and that, for all cost ratios above 2:1, models constructed with the extended HGA-NN technique give better results.

5.4. Results discussion and comparison

In most studies, the classifier performance is estimated by means of the predictive accuracy of the constructed models. The predictive accuracy is a less appropriate measure if the costs of different errors vary greatly[4] and if the classification aim is misclassification costs optimization. Furthermore, an array of measures provides a richer picture of model quality than only a single measure. Accordingly, in this study, the results were measured by using: accuracy, AUC, F-measure, F-β measure and the relative cost of misclassification. The last one is used as the main classification algorithm quality measure because our goal is to optimize the misclassification cost and generating the higher profit, and return on assets.

Therefore, the analysis of accuracy shows that the standard HGA-NN technique gives better results than the new extended HGA-NN technique, regardless of the extension used. The results were consistent in both datasets. Except accuracy, all experimental results obtained using the extended HGA-NN classifier are superior to those obtained using the standard HGA-NN classifier.

From Tables 4 and 6, as well as from Figures 2 and 3, which show the comparison of the relative misclassification costs of the HGA-NN and new extended HGA-NN techniques, it is clear that the resampling techniques, combined with the feature selection technique, contribute positively to the results in reducing the cost of misclassification. This contribution is more significant when the relative cost ratio is higher. Therefore, taken together, the HGA-NN ROS technique gives the best results. For the results comparison of the presented techniques with results presented in the literature for the German credit dataset, the results of the HGA-NN and HGA-NN ROS techniques will be used in accordance with Tables 5 and 6. The HGA-NN technique gives the most accurate prediction and the HGA-NN ROS technique results are the best for all other performance measures.

Table 7. The classification accuracy comparison of the HGA-NN and HGA-NN ROS techniques with results from the literature on the German dataset

Technique (algorithm)		Error probability		Accuracy
Name	Code	P_1 (%)[a]	P_2 (%)[b]	Mean (%)
HGA-NN ROS	Alg1	11.67	38.43	69.60
HGA-NN	Alg2	46.33	10.71	**78.60**
SVM[27]	Alg3	37.00	18.00	77.00
LogR[28]	Alg4	50.66	11.69	76.62
Bagging/MLP[29]	Alg5	49.40	24.60	75.33
Logit[30]	Alg6	18.33	44.00	63.70
RBF KASNP[31]	Alg7	26.25	28.69	72.05
FA MLP[32]	Alg8	48.69	10.66	77.93
BoostingSVM[33]	Alg9	54.38	10.56	76.29
logR[23]	Alg10	43.00	13.86	77.40

[a] P_1 - Type I error

[b] P_2 - Type II error

Table 8. The relative misclassification costs comparison of the HGA-NN and HGA-NN ROS techniques with results from the literature on the German dataset

Technique	Cost ratio (C_I:C_{II})						
	1:1	2:1	3:1	4:1	5:1	8:1	10:1
Alg1	0.3040	**0.3390**[a]	**0.3740**[a]	**0.4091**[a]	**0.4441**[a]	**0.5491**[a]	**0.6191**[a]
Alg2	**0.2140**[a]	0.3530	0.4919	0.6309	0.7699	1.1869	1.4649
Alg3	0.2370	0.3480	0.4590	0.5700	0.6810	1.0140	1.2360
Alg4	0.2338	0.3858	0.5378	0.6898	0.8417	1.2977	1.6016
Alg5	0.3204	0.4686	0.6168	0.7650	0.9132	1.3578	1.6542
Alg6	0.3630	0.4180	0.4730	0.5280	0.5830	0.7479	0.8579
Alg7	0.2796	0.3583	0.4371	0.5158	0.5946	0.8308	0.9883
Alg8	0.2207	0.3668	0.5128	0.6589	0.8050	1.2432	1.5353
Alg9	0.2371	0.4002	0.5633	0.7265	0.8896	1.3790	1.7053
Alg10	0.2260	0.3550	0.4840	0.6130	0.7420	1.1290	1.3870

[a] The best result for each cost ratio.

Many authors, including Tsai and Cheng [26], indicate that the German credit dataset is a challenging benchmark for bankruptcy prediction. Therefore, for a reliable and effective examination of the performance of the prediction models in the area of credit risk, one should consider the German dataset to be a benchmark for evaluation. For that reason and because it is a very limited number of publicly available high dimensional retail credit datasets, the retail credit dataset used in this experiment is the German credit dataset.

From the comparison of the results presented in the literature and the results of the HGA-NN and the HGA-NN ROS technique (Table 7), it can be observed that the HGA-NN technique achieves excellent results in respect to classification accuracy and 10-fold cross-validation. The comparison of the relative costs of misclassification (Table 8) shows that the new HGA-NN ROS technique achieves the best results. This is a result of: (a) the core algorithm, by which the technique makes simultaneous adjustment of: (1) data to the classification algorithm (feature selection) and (2) algorithm parameters to the data; and (b) the further extension which mitigates the negative impact of a class and cost imbalance to the cost efficiency. This is of particular importance in the domain for which the technique is developed and justifies additional efforts in its implementation.

Table 9. Relative costs of misclassification rankings for the German dataset

Technique	Cost ratio (C_I:C_{II})							sum of	average	R^2	Ne
	1:1	2:1	3:1	4:1	5:1	8:1	10:1	ranks	R		men
Alg1	8	1	1	1	1	1	1	**14**	2.000	4.000	
Alg2	1	3	6	6	6	6	6	34	4.857	23.592	
Alg3	5	2	3	4	4	4	4	26	3.714	13.796	
Alg4	4	7	8	8	8	8	8	51	7.286	53.082	[a]
Alg5	9	10	10	10	10	9	9	67	9.571	91.612	[a]
Alg6	10	9	4	3	2	2	2	32	4.571	20.898	
Alg7	7	5	2	2	3	3	3	25	3.571	12.755	
Alg8	2	6	7	7	7	7	7	43	6.143	37.735	
Alg9	6	8	9	9	9	10	10	61	8.714	75.939	[a]
Alg10	3	4	5	5	5	5	5	32	4.571	20.898	

[a] A significant difference compared to Alg1, with $\alpha = 0.05$.

The significance of result differences, measured by using the relative costs of misclassification, between the HGA-NN ROS techniques and the results presented in the literature on the German dataset was further tested by using the Friedman test. Therefore, in Table 9, the results of all algorithms are ranked for each cost ratio separately; the algorithm with the best score receives rank 1, the one with the second-best score receives rank 2, and so on. Substituting the values from Table 9 into Eq. (8), we get $\chi_F^2 = 39.561$.

The results obtained using the statistical tool R version 3.1.0 is: Friedman chi-squared = 39.561, df = 9, and p-value <0.0001. Because the p-value is significant ($\alpha = 0.05$), the Nemenyi test was applied. According to the Nemenyi post hoc test, the performance of two classifiers is significantly different if their average ranks differ by at least a critical difference. Substituting the parameters into Eq. (9), we obtain the critical difference CD = 5.12 with $\alpha = 0.05$.

From Table 9, it is clear that the HGA-NN ROS algorithm, i.e., the one with the lowest Friedman rank, according to the Nemenyi post hoc test, performs significantly better than the following algorithms: Alg5, Alg9 and Alg4. According to the Holm post hoc test, Alg1 is significantly better than the mentioned algorithms, and Alg8, on the level of significance for $\alpha = 0.05$. All experimental results confirm our hypothesis that resampling techniques combined with the feature selection technique have positive impact on the relative cost of misclassification in retail credit risk assessment.

6. Conclusions

This paper explored the impact of resampling techniques combined with the feature selection technique on the classification results in the domain of retail credit risk assessment. The research was performed on two datasets, i.e., Croatian and German. Classification performances of the extended HGA-NN technique were measured by using different measures of performance: accuracy, AUC, F-measure, F-β measure and the relative cost of misclassification. The focus was on the costs of misclassification. To assess the overall quality of the techniques presented here, an empirical comparison was performed with the results of other techniques reported in the literature. The Friedman test and Nemenyi post hoc test were applied to determine whether the result differences are statistically significant.

The results show that, with respect to average misclassification costs, the HGA-NN ROS technique achieves better results than other techniques presented in the literature for the cost ratio 2:1 and above. The Friedman test and Nemenyi post hoc test confirmed the results on the German dataset. As small differences in model power can lead to significant economic impact for the user, the results demonstrate the potential of the new technique for dealing with credit risk cost-sensitive imbalanced data in terms of misclassification costs. Accordingly, the results justify the added modelling complexity, especially when the banks have a high level of risk aversion. Overall, if the bank optimizes its loan approval process for cost ratios above 2:1, the results of this research indicate that some type of cost-sensitive learning can be very helpful.

Possible limitations of the study come from the fact that we conducted research on two datasets. This is the consequence of its focus to retail credit risk assessment. In this area, there is a very limited number of publicly available high dimensional datasets. Therefore, this limitation is difficult to overcome, but anyway, a research is needed. In addition, during this study, we recognized the need for a more thorough exploration of the relationship between type I and II errors in credit risk assessment because cost optimization depends significantly on this relationship.

References

[1] Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 875-886). Springer, Boston, MA.

[2] Naganjaneyulu, S., & Kuppa, M. R. (2013). A novel framework for class imbalance learning using intelligent under-sampling. *Progress in Artificial Intelligence*, *2*(1), 73-84.

[3] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113-141.

[4] 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*(2-3), 427-436.

[5] 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.

[6] Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, *32*(8), 1541-1551.

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

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

[9] Oreški, G., & Oreški, S. (2015). Two Stage Comparison of Classifier Performances for Highly Imbalanced Datasets. *Journal of Information and Organizational Sciences*, *39*(2), 209-222.

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

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

[12] 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.

[13] 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.

[14] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

[15] Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2017). Cost-Sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*.

[16] Braytee, A., Liu, W., & Kennedy, P. (2016, October). A cost-sensitive learning strategy for feature extraction from imbalanced data. In *International Conference on Neural Information Processing* (pp. 78-86). Springer, Cham.

[17] Cieslak, D. A., & Chawla, N. V. (2008, September). Learning decision trees for unbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 241-256). Springer, Berlin, Heidelberg.

[18] Han J. and Kamber M. (2006). *Data Mining: Concepts and Techniques*, Second Edi. 500 Sansome Street, Suite 400, San Francisco, CA 94111: Morgan Kaufmann Publishers.

[19] Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, *41*, 16-25.

[20] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[21] 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.

[22] López, V., Fernández, A., & Herrera, F. (2014). On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, *257*, 1-13.

[23] Marqués, A. I., García, V., & Sánchez, J. S. (2012). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*, *39*(12), 10916-10922.

[24] Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

[25] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, *7*(Jan), 1-30.

[26] Tsai, C. F., & Cheng, K. C. (2012). Simple instance selection for bankruptcy prediction. *Knowledge-Based Systems*, *27*, 333-342.

[27] Han, L., Han, L., & Zhao, H. (2013). Orthogonal support vector machine for credit scoring. *Engineering Applications of Artificial Intelligence*, *26*(2), 848-862.

[28] Zhu, X., Li, J., Wu, D., Wang, H., & Liang, C. (2013). Balancing accuracy, complexity and interpretability in consumer credit decision making: A C-TOPSIS classification approach. *Knowledge-Based Systems*, *52*, 258-267.

[29] Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, *36*(2), 3028-3033.

[30] Lieli, R. P., & White, H. (2010). The construction of empirical credit scoring rules based on maximization principles. *Journal of Econometrics*, *157*(1), 110-119.

[31] Zhou, X., Jiang, W., Shi, Y., & Tian, Y. (2011). Credit risk evaluation with kernel-based affine subspace nearest points learning method. *Expert Systems with Applications*, *38*(4), 4272-4279.

[32] Tsai, C. F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, *22*(2), 120-127.

[33] Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, *38*(1), 223-230.