Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction

Vicente García[a], Ana I. Marqués[b], J. Salvador Sánchez[c,*]

[a] *Multidisciplinary University Division, Universidad Autónoma de Ciudad Juárez Ciudad Juárez, Chihuahua, Mexico*
[b] *Department of Business Administration and Marketing, Universitat Jaume I Castelló de la Plana, Spain*
[c] *Institute of New Imaging Technologies, Department of Computer Languages and Systems, Universitat Jaume I, Castelló de la Plana, Spain*

Abstract

Credit risk and corporate bankruptcy prediction has widely been studied as a binary classification problem using both advanced statistical and machine learning models. Ensembles of classifiers have demonstrated their effectiveness for various applications in finance using data sets that are often characterized by imperfections such as irrelevant features, skewed classes, data set shift, and missing and noisy data. However, there are other corruptions in the data that might hinder the prediction performance mainly on the default or bankrupt (positive) cases, where the misclassification costs are typically much higher than those associated to the non-default or non-bankrupt (negative) class. Here we characterize the complexity of 14 real-life financial databases based on the different types of positive samples. The objective is to gain some insight into the potential links between the performance of classifier ensembles (BAGGING, AdaBoost, random subspace, DECORATE, rotation forest, random forest, and stochastic gradient boosting) and the positive sample types. Experimental results reveal that the performance of the ensembles indeed depends on the prevalent type of positive samples.

Keywords: Types of samples, Credit risk, Bankruptcy, Classifier ensemble,

*Corresponding author. Tel.: +34 964 728 350
Email addresses: vicente.jimenez@uacj.mx (Vicente García), imarques@uji.es (Ana I. Marqués), sanchez@uji.es (J. Salvador Sánchez)

Imbalance.

1. Introduction

In response to the 2008 global financial crisis, banks and regulatory agencies have increased their efforts to streamline processes and increase efficiency in the prediction and proactive management of credit risk, financial distress and corporate bankruptcy. Classical studies on this subject were initially based on advanced statistical models [1, 2, 3, 4, 5], such as logistic regression, probit analysis, linear discriminant analysis, survival analysis, linear and quadratic programming, and multivariate adaptive regression splines. Nevertheless, empirical results have shown that most underlying assumptions of these statistical approaches, such as multivariate normality and independence of the explanatory variables, are frequently violated [6, 7].

Unlike the statistical models, machine learning and computational intelligence methods do not assume any specific prior knowledge, but instead they automatically extract information from past observations. These are represented by a set of explanatory variables, which usually correspond to financial ratios, macroeconomic indicators and socio-demographic characteristics, either straightforwardly represented as continuous variables or discretized as qualitative information.

Support vector machines [8, 9, 10], genetic and evolutionary algorithms [11, 12, 13], artificial neural networks [14, 15, 16, 17, 18], rough sets [19, 20, 21], and decision trees [22, 23] have received much attention and widespread application in the field of finance and more specifically, to the prediction of credit risk, financial distress and corporate bankruptcy. Although numerous previous studies concluded that machine learning techniques are superior to statistical models, it has been argued that no single classifier can produce the best results on all the cases. From this conclusion, ensembles emerged as a powerful tool for exploiting the different behavior of a pool of individual (base) learners and reducing prediction errors in several financial applications. In fact, practical investigations have demonstrated that ensembles generally outperform stand-alone prediction methods in most credit risk and corporate bankruptcy prediction problems [24, 25, 26, 27]. However, extensive researches have also shown the strengths and weaknesses of classifier ensembles against a diversity of intrinsic data characteristics, which could make the prediction of the positive cases even much more difficult; for in-

stance, one can find studies on class imbalance [28], attribute noise [29], and data set shift [30], among others.

Since the error rate of default or bankrupt (positive) cases is of great importance for credit risk and corporate bankruptcy assessment, it could be useful to carry out a proper analysis on how the presence of samples of different nature in the positive class may affect the predictive performance of classifier ensembles. However, as far as we are aware, no previously reported study has systematically analyzed this problem in the framework of finance.

Therefore, considering the particular characteristics of financial data, the ultimate aim of this paper is to characterize the databases according to the prevalent type of samples in the minority class and also to explore the potential links between the performance of classifier ensembles and the different types of data sets. To this end, experiments will consist of characterizing 14 credit and bankruptcy data sets according to the positive sample types, and analyzing whether or not there may exist any correlation between these and the performance of several prediction systems based upon seven well-established ensembles that are built with three different base classifiers. As the number of positive samples is usually far less than the amount of negative samples, which leads to the well-renowned class imbalance problem, we cannot neglect this scenario when discussing the experimental results.

Henceforth the rest of the paper is organized as follows. Section 2 reviews some research works related to the use of ensembles to deal with various intrinsic data characteristics in the field of credit risk and corporate bankruptcy prediction. Next, Section 3 provides a categorization of the types of samples that can be found in a data set. The experimental set-up, databases and classifier ensembles are given in Section 4, whereas the results are reported and discussed in Section 5. Finally, the conclusions and possible avenues for further research are outlined in Section 6.

2. Intrinsic financial data characteristics in ensembles

The development of classifier ensembles for credit risk, financial distress and corporate bankruptcy prediction has attracted increasing attention of both researchers and practitioners in the last years. Many works have shown the superiority of ensembles over single classifiers, whereas some others have proposed new algorithms as alternatives to the existing ones. However, only a few works have paid attention to studying the behavior of ensembles when learning from data sets with several intrinsic data characteristics. Here, we

summarize some of the most recent publications on this topic, but note that it does not intend to be a thorough review.

Das et al. [31] proposed that intrinsic data characteristics can be categorized into two groups: 1) distribution-based data irregularity (DistBI), and 2) feature-based data irregularity (FeatBI). The former involves class imbalance problem, small disjuncts, and class distribution skew, and the latter includes missing and absent features. We consider that the first group might also cover other problems, such as outliers, noisy data, small data set size, and data set shift. Analogously, we believe that the second group might also include noisy, irrelevant and redundant features. Taking this taxonomy into account, the Venn diagram in Figure 1 shows the relationship between both categories and the number of works in each group (see Table A.4 of Appendix A for a more detailed information). As can be seen, a majority of works have focused on the distribution-based data irregularities, where the class imbalance appears as the most studied problem. Only three works have faced the feature-based data irregularities, whereas five of them addressed both intrinsic data characteristics.

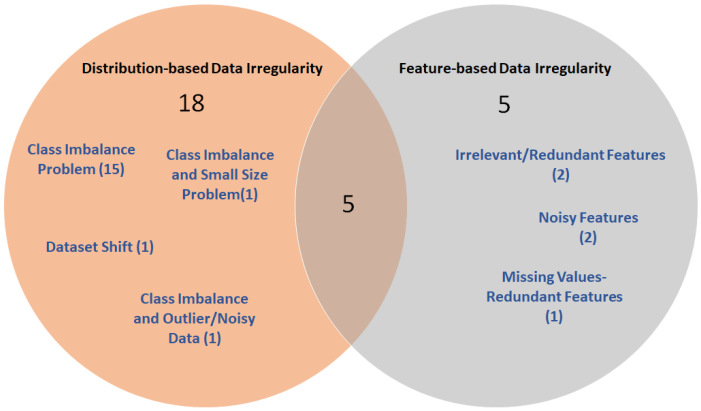


Figure 1: A Venn diagram of the intrinsic financial data characteristics

2.1. Distribution-based data irregularity

The class imbalance problem has been considered as a challenging task in a broad scope of financial problems. In last years, it has been very frequent to deal with imbalanced data, and several works [32, 33] have studied the

performance of many different ensembles on data sets with this intrinsic data characteristic.

Feng et al. [34] presented a dynamic ensemble model based on soft probability where the classifier selection was based on accuracy, precision and different costs of type I error and type II error. Experimental results showed that the proposed model outperforms BAGGING and random forest on several imbalanced credit data sets. In the same line, Xiao et al. [35] combined the dynamic classifier selection method with a cost-sensitive evaluation criteria. He et al. [36] introduced a cascade model that resamples the credit scoring data sets according to their imbalance ratio and a threshold. Each adjusted data set is used for training several random forests and extreme gradient boosting as base classifiers. Sun et al. [37] proposed an ensemble for imbalanced credit evaluation based on the SMOTE algorithm and the BAGGING technique with different sampling rates. Wang et al. [38] combined the Lasso-logistic regression model with the BAGGING approach where this was used on the minority class to generate balanced training data sets. Sun et al. [39] combined SMOTE with the BAGGING algorithm using a support vector machine (SVM) as base learner. While all these works are characterized by incorporating the imbalance solutions into the ensemble, Louzada et al. [40] developed a new BAGGING algorithm, called Poly-BAGGING, where the resampling technique was not considered as part of the ensemble approach.

Xia et al. [41] designed a heterogeneous ensemble credit scoring model by integrating the BAGGING algorithm with the stacking method; despite the model introduced did not focus on class imbalance problems, it showed a good performance on moderately imbalanced data sets. Yu et al. [42] developed a three-stage ensemble model for dealing with class imbalance problems using BAGGING, SVM and a deep belief network. Abellán and Castellano [33] showed that an ensemble built with the credal decision tree performs better than others based on more complex base learners trained on balanced and imbalanced data sets. Ala'raj and Abbod [43] introduced a new combination approach based on classifier consensus that creates a ranking group as a fusion of individual classifiers. Experimental results showed that the consensus model achieves better performance in terms of the H-measure on highly imbalanced data sets. Florez-Lopez and Ramon-Jeronimo [44] developed a novel ensemble technique that follows a three-stage structured called the correlated-adjusted decision forest. Empirical results revealed their suitability on imbalance problems in terms of type I error and type II error. Kim et al. [45] proposed the geometric mean based boosting algorithm, which is a

modification of AdaBoost using the concept of geometric error and accuracy calculation. Ziba et al. [46] used the extreme gradient boosting where each base learner was constructed using synthetic random features; the aim was to deal with class imbalance and small size problems. Li et al. [47] proposed a three-stage ensemble framework where in the first level, several perceptrons were used as base learners. In the middle level, a relevance vector machine was used to train weak learners. In the top of the framework a boosting algorithm was employed. Authors suggested that their proposal is suitable when the data set is imbalanced and there are noisy data. The data set shift problem occurs when the training and test data come from different distributions. To deal with this problem, Xiao et al. [30] proposed to use transfer learning into an ensemble model.

2.2. Feature-based data irregularity

Twala [48] performed an analysis on the behavior of several ensemble models when the data set shows different levels of attribute noise. The experimental results suggested that the impact of noise depends upon the classifier and the proportion of noise.

To eliminate irrelevant and redundant features, Muslim et al. [49] combined split feature reduction and BAGGING. Xia et al. [50] introduced a sequential extreme gradient boosting model that incorporates a preprocessing step to scale the data and handle missing values. In addition, a feature selection system was used to remove redundant variables. Koutanaei et al. [51] used feature selection algorithms as a first stage to remove noisy attributes. The reduced data sets were used on AdaBoost, BAGGING, random forest, and stacking. Wang et al. [52] introduced a feature selection algorithm into boosting to deal with irrelevant features.

2.3. Intersection between DistBI and FeatBI

Each intrinsic data characteristic does not constitute an isolated problem. Ala'raj and Abbod [53] used two preprocessing techniques, Gabriel neighborhood graph editing and multivariate adaptive regression splines, to reduce the size of the data set by filtering samples and choosing the most relevant features. Both algorithms were combined with a consensus ranking approach. Liao et al. [54] introduced an ensemble model with majority vote that combines SVM, multiple feature selection, artificial neural network (ANN), and rough set theory (RST). The SVM model was used to balancing the training set followed by a multiple feature selection algorithm to pick

up the most representative features. To deal with noisy data and the class imbalance problem, Li et al. [47] proposed a relevance vector machine ensemble model that employs a soft margin boosting. Wang et al. [55] introduced a two-stag ensemble model based on decision tree, BAGGING and random subspace to deal with the noise data and redundant attributes. Paleologo et al. [56] proposed a sub-BAGGING algorithm where the base learners were generated by random sub-sampling in order to handle the class imbalance problem. Besides, an imputation method integrated into the ensemble model was used to handle missing data.

3. Types of samples

When analyzing the characteristics of a data set, an important question that deserves to pay some special attention refers to the identification of the different types of samples. This identification can be particularly useful to support interpretations of differences in the performance of classifiers because many data complexity factors are linked to the distribution of sample types in a data set [57, 58].

According to the categorization proposed by several authors, two main types of samples can be distinguished: safe and unsafe [59, 60, 61]. Safe samples refer to those placed in homogeneous regions with data of a single class and are sufficiently separated from examples of the other class, whereas the remaining samples are deemed as unsafe. Most models classify the safe samples correctly, but the unsafe samples may make their learning especially difficult and more likely to be misclassified.

The property common to the unsafe samples is that they are located close to examples that belong to some different class. However, this type of samples can be further divided into three subgroups depending on their particular characteristics: borderline, rare and outlier [60, 62]. Borderline samples are located near the decision boundary between classes. Rare samples are small groups of examples located far from the core of their class, creating small data chunks or sub-clusters. Finally, the outliers are single samples that are surrounded by examples from the other class.

A simple method to identify each sample type is based on analyzing the local neighborhood of the examples [60, 61], which can be modeled either by their k-neighborhood or by using a kernel function. Thus, a safe sample is characterized by having a neighborhood dominated by examples that belong

to its same class. Rare examples and outliers are mainly surrounded by examples from different classes, whereas the borderline samples are surrounded by examples both from their same class and also from a different class.

Following the standard strategy used in prior works [58, 60, 61, 63], we determine the type of a sample s by comparing the number of its k nearest neighbors (with a constant value of $k = 5$) that belong to the class of s with the number of neighbors from the opposite class. Most authors choose $k = 5$ because smaller values may poorly distinguish the nature of examples and higher values would violate the local neighborhood assumption. Thus we can find the following cases:

- A sample s is considered to be safe if at least 4 out of the 5 nearest neighbors belong to the class of s.
- A sample s is considered to be borderline if 2–3 out of its 5 nearest neighbors belong to the class of s.
- A sample s is considered to be rare if only one nearest neighbor belongs to the class of s, and this has no more than one neighbor from its same class.
- A sample s is considered to be outlier if all its nearest neighbors are from the opposite class.

This method has been proposed for the identification of the different sample types in the minority class, which is especially relevant when the class distribution is imbalanced. Note that in such a situation, the percentage of each sample type belonging to the majority and minority classes may differ massively from each other. For instance, consider a credit data set where only 1% of samples are defaulters and 99% are non-defaulters; under these conditions, it is likely that most of the safe samples belong to the majority class and most of the unsafe samples are in the minority class, which may disguise the true distribution of sample types in the data set.

4. Databases and experimental set-up

The experiments were designed to explore the potential impact of the different sample types on the prediction performance of classifier ensembles over a collection of bankruptcy and creditworthiness data sets. Table 1 summarizes the main characteristics of the databases, reporting the number of

explanatory variables, the amount of positive and negative examples, the total number of cases, and the imbalance ratio (IR) defined as the ratio of the number of negative examples to the number of positive examples. Data sets with an IR ≥ 10 have been defined as strongly imbalanced. All databases represent two-class problems with different levels of imbalance, which ranges from 0.80 in Australian (i.e., the class of most interest outnumbers the other class) to 24.93 in Polish-1st.

Table 1: Characteristics of the data sets

		#Variables	#Positive	#Negative	#Samples	IR	Source
Low imbalance	Australian	14	383	307	690	0.80	[64]
	Finland	40	250	250	500	1.00	[65]
	SabiSPQ	16	472	472	944	1.00	[26]
	Polish	30	112	128	240	1.14	[66]
	Japanese	15	296	357	653	1.21	[64]
	German	24	300	700	1000	2.33	[64]
	Thomas	12	323	902	1225	2.79	[67]
	Taiwan	23	6636	23364	30000	3.52	[68]
High imbalance	Polish-5th	64	410	5500	5910	13.41	[64]
	Polish-4th	64	515	9277	9792	18.01	[64]
	Iranian	27	50	950	1000	19.00	[69]
	Polish-3rd	64	495	10008	10503	20.22	[64]
	Polish-2nd	64	400	9773	10173	24.43	[64]
	Polish-1st	64	271	6756	7027	24.93	[64]

A 10-fold cross-validation procedure was adopted with the purpose of avoiding biased results [70]. Each data set was randomly split into ten stratified blocks (or folds) of equal size. For each round, nine blocks are used for training and the remaining part for testing. This is repeated ten times using a different block for testing, thus ensuring that all folds are employed for both training and testing.

The performance of the classifiers was evaluated with three standard scores that have typically been used in financial applications. First, the area under the ROC curve (AUC) corresponds to an overall performance measure that allows decision-makers to compare examples against each other. Second, the true-positive rate (TPR) and true-negative rate (TNR) exhibit performance results on each class separately, thus taking care of the cost of different error types. Both these are particularly meaningful for the kind of real-life applications faced in this paper because the cost of false-negatives (predicting a default or bankrupt case as non-default or non-bankrupt) is often much higher than the cost associated to false-positives (non-defaulters predicted as defaulters) [71].

4.1. Ensembles of classifiers

We have chosen seven standard ensembles to evaluate whether or not there exists any connection between their prediction performance and the sample types in the data sets: BAGGING (*Bootstrap AGGregatING*, Bag), AdaBoost (ABoost), stochastic gradient boosting (SGBoost), random subspace (RSP), rotation forest (RotF), random forest (RndF), and DECORATE (*Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples*, Decor).

The Bag technique [72] generates multiple bootstrap samples randomly drawn with replacement from the original training set. Next, each individual classifier is built for each sample, and predictions on new cases are made by combining the classification results using a majority voting policy.

Boosting [73] produces a sequence of base classifiers through successive bootstrap samples that are obtained by weighting the training data in a number of iterations. Initially equal weights are assigned to all training examples and at each iteration, boosting increases the weights on the examples predicted incorrectly by the previous individual classifier so that those misclassified examples are more likely to be chosen in the next bootstrap sample. Final decisions are based on a weighted majority voting scheme. Two of the most popular boosting algorithms are ABoost and SGBoost [74] (at each iteration a subsample of the training data is drawn at random without replacement from the full training set; the randomly selected subsample is then used, instead of the full sample, to fit the base learner).

In the RSP method proposed by Ho [75], the base classifiers are trained on sets constructed with a given proportion of variables picked randomly from the original set of features. The outputs of the individual classifiers are then combined into a final decision rule through a simple majority voting procedure.

RotF [76] trains each base classifier with a different set of extracted variables. The original feature set is randomly split into a number of subsets, principal component analysis is run separately on each subset, and a new set of linear extracted variables is constructed by pooling all principal components. The data is transformed linearly into the new feature space, and the base classifier is trained with this new data set.

The RndF developed by Breiman [77] is an ensemble of decision trees, each one built using a bootstrap sample of the training data and the candidate set of variables at each split is a random subset of the features. Each tree is

unpruned, so as to obtain low-bias trees; in addition, bagging and random variable selection result in low correlation of the individual trees.

The Decor algorithm [78] uses a base learner to build an ensemble iteratively by adding different randomly generated examples to the training set when building new ensemble members. These artificially generated examples are given class labels that disagree with the prediction of the current ensemble, thereby increasing diversity when a new classifier is trained on the augmented data and incorporated into the ensemble.

The ensembles were built using different base classifiers that have been widely used in the financial industry: the unpruned C4.5 decision tree, the multi-layer perceptron (MLP) with one hidden layer and the k nearest neighbors (kNN) rule. These base classifiers have been chosen because it is known that some ensembles such as bagging do not work well with stable (low variance and possibly high bias) models. The hyperparameters for the MLP were tuned by withholding 50% of the training data as a validation set and testing the learning rate and the momentum from 0.1 to 0.3. The kNN rule was optimized using leaving-one-out on the training set to select the best k value between 1 and 30.

The following ensemble configurations were investigated:

- BAGGING: Bag(MLP), Bag(C4.5), Bag(kNN).
- AdaBoost: ABoost(MLP), ABoost(C4.5), ABoost(kNN).
- Random subspace: RSP(MLP), RSP(C4.5), RSP(kNN).
- DECORATE: Decor(MLP), Decor(C4.5), Decor(kNN).
- Rotation forest: RotF(C4.5).
- Random forest: RndF(C4.5).
- Stochastic gradient boosting: SGBoost(C4.5).

4.2. Characterization of the databases

To gain a better insight into the structure of the classes and a deeper understanding of the data complexity, the data sets are here characterized according to the sample types introduced in Section 3 and the imbalance ratio reported in Table 1. This can make easier the subsequent analysis between

the predictive performance of classifier ensembles and the distribution of sample types in the data sets.

For each database, the percentage of samples that belong to each type was calculated for both the positive class (Figure 2) and the negative class (Figure 3). The sample types were displayed as bar charts on the left y-axis of these graphs and the imbalance ratio was displayed on the right y-axis as a series plot. The data sets on the x-axis were sorted in ascending order of the imbalance ratio.

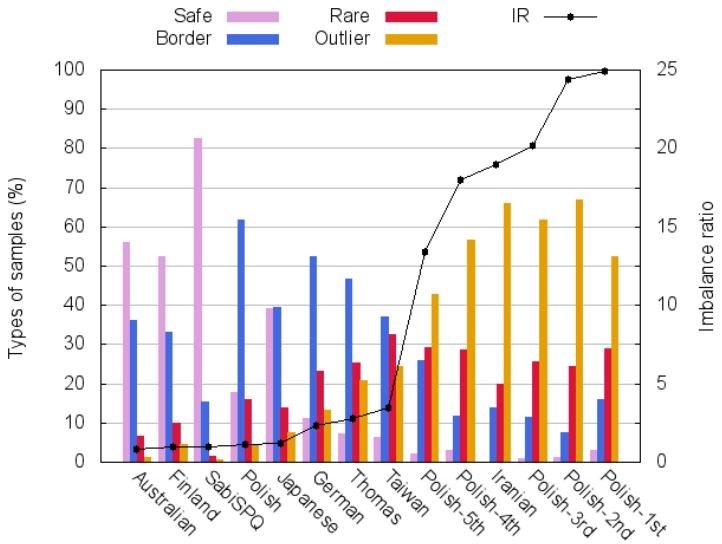


Figure 2: Percentages of sample types (left y-axis) and imbalance ratio (right y-axis) for the positive class

Not surprisingly, comparison of both figures shows that the distribution of sample types in each class strongly depends on the imbalance ratio. Thus the amount of safe positive samples in the data sets with high imbalance was minimal (less than 3%), whereas the number of rare and outlier samples in the positive class increased with the imbalance ratio. On the other hand, for the highly imbalanced databases the proportion of safe samples in the negative class was very close to 100% and the percentage of unsafe samples was nearly 0%: the maximum amount of borderline and rare samples was for the Polish-5th database (4.69% and 0.27%, respectively), and that of outliers was for the Polish-1st database (0.03%).

These findings reinforce the idea that performing an accurate prediction

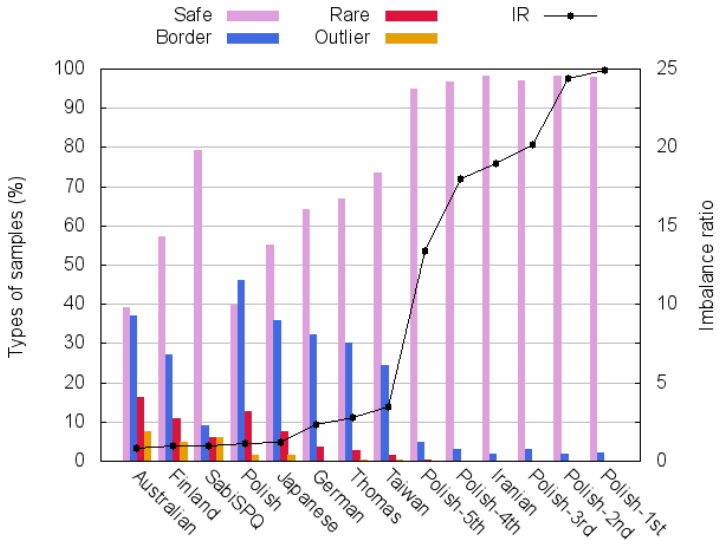


Figure 3: Percentages of sample types (left y-axis) and imbalance ratio (right y-axis) for the negative class

is more complicated in the positive class than in the negative class because their proportions of safe and unsafe samples are very different, especially in the data sets with high imbalance; notwithstanding, some cases deserve further comments. For instance, both Finland and SabiSPQ data sets are perfectly balanced (IR = 1.0), but the former has a lower amount of safe samples in the positive class than the latter (52.40% and 82.42%, respectively). Analogously, the Polish and Japanese databases, which are characterized by similar imbalance ratios, present very different percentages of safe samples in the positive class (17.86% and 39.19%, respectively). Even more interesting is the comparison between SabiSPQ (IR = 1.0) and Polish (IR = 1.14) because there exist large differences in the amount of safe and borderline samples in the minority class of each database. This suggests that, as pointed out in other research works [57, 63, 79], the class imbalance is not the only problem that may degrade the classifier performance, but there are other intrinsic data characteristics that also hinder classification; therefore, an analysis of the overall structure of data can become of great relevance because it could provide some insights on choosing the most appropriate classifier ensemble depending on the distribution of sample types.

From Figure 2, it is possible to categorize the experimental databases into

five groups according to the prevalent type of samples in the minority class:

- Safe: The Australian, Finland and SabiSPQ databases, which contain more than 50% of safe samples.
- Borderline: The Polish, German and Thomas databases, which have a majority of borderline samples (45%-60%).
- Outlier: All the high imbalanced data sets because more than 40% of examples have been characterized as outliers.
- Safe-borderline: The Japanese database has approximately the same amount of safe and borderline examples in the positive class, about 40% each one.
- Borderline-rare: The Taiwan database comprises a majority of borderline and rare samples, which represent close to 70% of the minority class: 37% of borderline samples and 32% of rare samples.

The last four groups refer to unsafe databases because less than 50% of their positive samples have been identified as safe. On the other hand, the last two categories correspond to data sets in which the positive samples are mainly placed between the safe and the borderline groups in the first case or between the borderline and the rare groups in the second one.

5. Results and discussion

We report the results obtained in the course of the experimental study. The aim is to investigate how the prevalent type of positive samples affects the performance of each ensemble model. In other words, the question to answer here is whether or not there exists any difference in performance of the classifier ensembles from one category of databases to another.

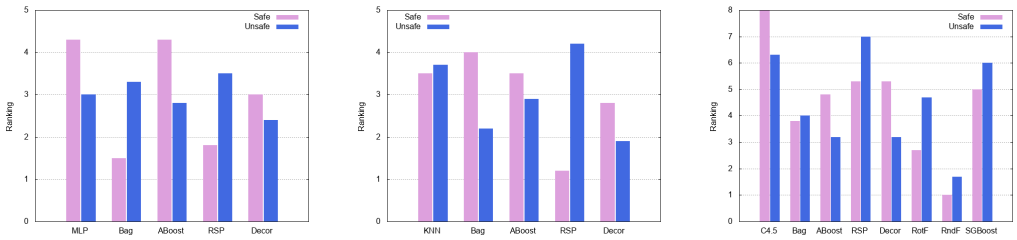
We have divided this section into two parts. First, we investigate the performance of the ensembles by comparing the safe databases against all the unsafe data sets (i.e., safe-borderline, borderline, borderline-rare and outlier data). The second part of this section is devoted to analyze the behavior of the ensembles for each category of unsafe databases.

As the values of AUC, TPR and TNR can be very different from one data set to another, the use of the average scores across the databases could be

inadequate. Instead we calculated Friedman's average ranks for the classifier ensembles. From this, one has to consider that the prediction model with the lowest average rank corresponds to the best algorithm. In addition, the full set of results is provided in Tables B.5–B.10 of Appendix B.

5.1. Performance analysis of the ensembles on the safe and unsafe data

The focus of the first block of experiments is on analyzing the possible differences in the behavior of the ensembles between the safe databases and the unsafe databases. To this end, Figure 4 displays the Friedman's average ranks of AUC for each classifier ensemble applied to both these categories of data sets. The stand-alone classifiers were also included in these plots as a baseline. As can be observed, there exist significant differences between the safe data and the unsafe data, irrespective of the base classifier used to build the ensembles.



(a) MLP (b) kNN (c) C4.5

Figure 4: Average ranks of AUC for safe and unsafe databases

In the case of the MLP-based models, the best ensembles were bagging and random subspace for the safe databases, and DECORATE and AdaBoost for the unsafe data sets. It is particularly appealing the behavior of AdaBoost because this was one of the best ensembles on the unsafe databases, but it performed even worse than the stand-alone MLP on the safe databases. Similar comments can be made for the kNN-based ensembles, in which RSP obtained the lowest average rank for the safe data sets and it was the worst technique for the unsafe ones. For the C4.5-based ensembles, the random forest was the best performing method for both safe and unsafe data, but the remaining models showed a significantly different behavior when applied to safe or unsafe data sets.

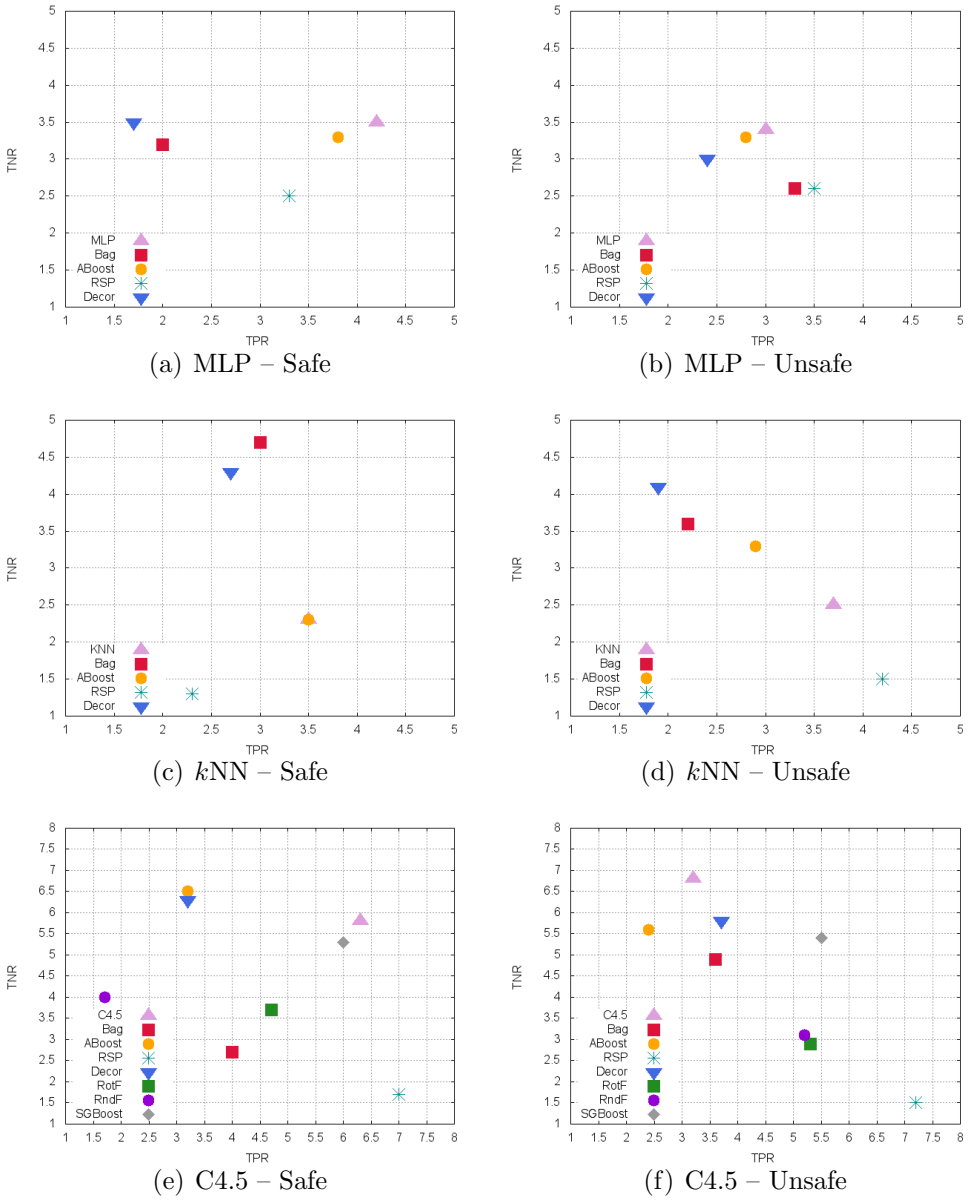
As AUC represents an overall, scalar performance evaluation measure, it can give rise to misleading conclusions when the cost of misclassifying examples in one class is very different from the cost of misclassifying examples in the other class, or when the class distribution is imbalanced [80, 81, 82]. In such cases, it is also especially important to evaluate the true-positive and true-negative rates. The former is the primary goal in credit risk and corporate bankruptcy prediction, but high true-positive rates should not compromise the correct classification of the majority class. To balance these two competing goals, a normalized Euclidean distance between each (average rank of TPR, average rank of TNR) pair and the origin (1, 1) was calculated and reported in Table 2. Using this measure, the best model for each type of data was the one that produced the smallest distance (highlighted in bold in Table 2).

Table 2: Distance measure for the safe and unsafe databases

	Safe	Unsafe
MLP	3.504	3.160
Bag(MLP)	**2.386**	2.838
ABoost(MLP)	3.670	2.918
RSP(MLP)	2.774	2.950
Decor(MLP)	2.587	**2.447**
kNN	2.833	3.135
Bag(kNN)	4.177	**2.867**
ABoost(kNN)	2.833	2.968
RSP(kNN)	**1.374**	3.266
Decor(kNN)	3.727	3.222
C4.5	7.197	6.214
Bag(C4.5)	3.432	**4.677**
ABoost(C4.5)	5.911	4.833
RSP(C4.5)	6.037	6.206
Decor(C4.5)	5.757	5.475
RotF(C4.5)	4.534	**4.703**
RndF(C4.5)	**3.073**	**4.696**
SGBoost(C4.5)	6.616	6.236

Another way of visualizing this consisted of plotting the Friedman's average ranks of TPR versus the Friedman's average ranks of TNR in Figure 5, and looking for the point that was closest to the bottom left corner of the graphs; thus the closer the ensemble was to the bottom left corner, the higher the performance on both classes. Note that this graph depicts relative trade-offs between TPR and TNR.

Although the ultimate objective of any classification system is to achieve high rates on both classes (that is, the ensembles with the smallest distance



(a) MLP – Safe (b) MLP – Unsafe

(c) kNN – Safe (d) kNN – Unsafe

(e) C4.5 – Safe (f) C4.5 – Unsafe

Figure 5: Average ranks of TPR vs. average ranks of TNR for the safe and unsafe databases

in Table 2), in general it will be preferable to maximize TPR rather than

to maximize TNR. This means that the ensembles close to the left side of the charts will be considered better than the ensembles close to the bottom side. One can observe in Figure 5 that the best models were: (i) BAGGING for the safe databases and DECORATE for the unsafe ones in the MLP-based models; (ii) RSP for the safe data sets and BAGGING for the unsafe ones in the kNN-based ensembles; and (iii) random forest for the safe data and BAGGING, random forest and rotation forest (all three algorithms with nearly the same normalized distance) for the unsafe ones in the C4.5-based methods. As can be seen, some of these conclusions do not agree with those drawn from the AUC analysis because this measure can be biased towards the majority class.

In summary, the main conclusion from this first analysis is that there indeed exist differences in the performance of the ensembles depending on the prevalent type of data, that is, each particular classifier ensemble does not perform equally well on safe and unsafe databases. Therefore, the next step will be to explore the behavior of the ensembles on the different types of databases that belong to the general category of unsafe data in order to investigate the possible links between each type and the performance of the ensembles.

5.2. Performance analysis of the ensembles on the unsafe data

The purpose of the second block of experiments is to establish the best performing ensemble for each type of unsafe databases when using each of the base classifiers. Thus Table 3 reports the normalized distance measure for all ensembles and the four types of unsafe data. In addition, each graph in Figures 6, 7 and 8 displays the Friedman's average ranks of TPR against those of TNR given by the MLP-based, kNN-based and C4.5-based ensembles, respectively.

Focusing on the results of the MLP-based ensembles in Table 3 and Figure 6, one can observe that the DECORATE algorithm achieved the highest performance (the smallest distance) for the safe-borderline databases, which correspond to the easiest type of unsafe data. BAGGING was the best ensemble for the borderline-rare databases. In the case of the outlier data sets (i.e., the most complex data structures), both DECORATE and random subspace were the techniques with the smallest normalized distance. On the other hand, in general the performance of AdaBoost was similar to that of the stand-alone MLP classifier, thus suggesting that this ensemble configuration is of little value to deal with most types of unsafe data sets. In summary,

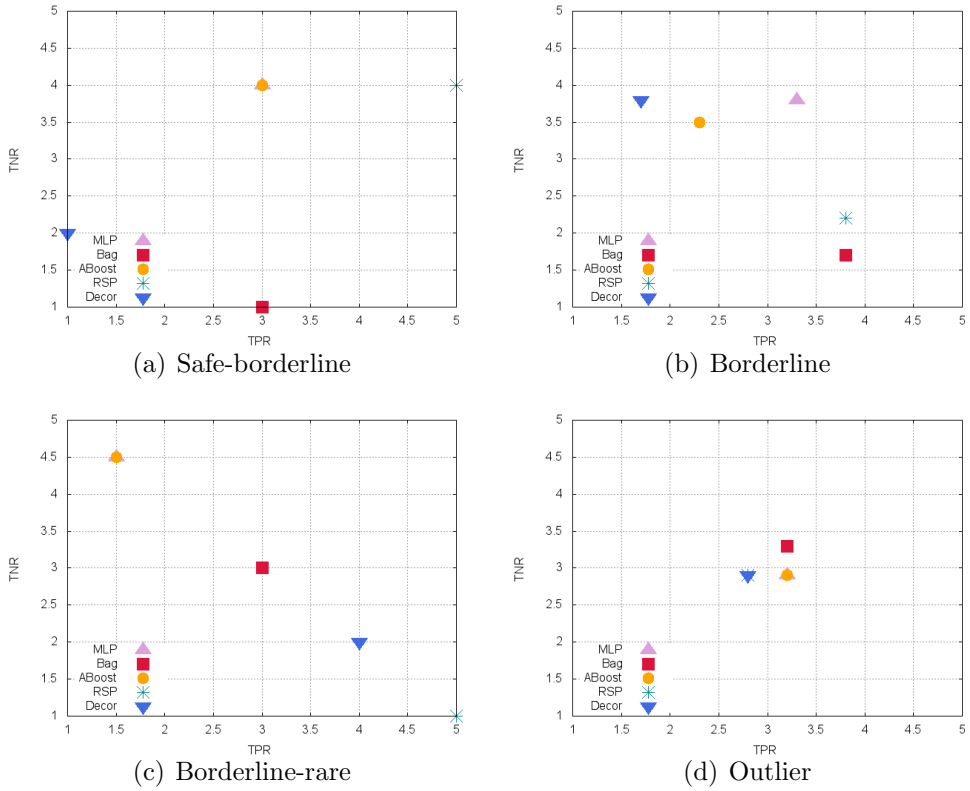
Table 3: Distance measure for the unsafe databases

	Safe-borderline	Borderline	Borderline-rare	Outlier
MLP	3.606	3.670	3.536	2.893
Bag(MLP)	2.000	2.911	**2.828**	3.184
ABoost(MLP)	3.606	**2.833**	3.536	2.893
RSP(MLP)	5.000	3.064	4.000	**2.595**
Decor(MLP)	**1.000**	2.911	3.162	**2.595**
kNN	3.536	3.606	3.162	3.111
Bag(kNN)	4.272	**2.386**	3.162	3.064
ABoost(kNN)	3.536	3.606	4.000	**2.801**
RSP(kNN)	**2.000**	2.877	4.000	3.563
Decor(kNN)	3.354	2.734	**2.828**	4.000
C4.5	6.325	6.379	7.211	6.692
Bag(C4.5)	6.021	5.918	4.000	**4.305**
ABoost(C4.5)	4.123	5.044	7.071	4.853
RSP(C4.5)	7.000	5.588	7.000	6.335
Decor(C4.5)	7.810	4.534	7.810	5.376
RotF(C4.5)	6.325	4.488	5.099	4.953
RndF(C4.5)	**3.041**	**3.745**	**2.828**	6.038
SGBoost(C4.5)	5.000	6.412	4.243	6.872

it seems that both DECORATE and BAGGING can be claimed as the best overall MLP-based ensembles when there is a majority of unsafe samples in the positive class.

With regards to the kNN-based methods, Table 3 indicates that RSP was the best performing algorithm for the safe-borderline data sets, BAGGING for the borderline data, DECORATE for the borderline-rare data, and AdaBoost for the outlier databases. However, the observation of plots in Figure 7 reveals that BAGGING achieved a lowest average rank of TPR than DECORATE for the borderline-rare data sets, thus suggesting that the former could be better than the latter. Similarly, BAGGING could also be considered to be better than AdaBoost for the outlier databases because it obtained a significantly lower average rank of TPR.

Finally, looking at the results obtained with the C4.5-based ensembles in Table 3, it is apparent that random forest was the most powerful model when the databases were characterized by a majority of safe-borderline, borderline or borderline-rare samples in the positive class. For these databases, Figure 8 shows that other ensembles achieved the best average ranks of TPR, but at the cost of producing very significantly higher error rates on the negative class; for instance, AdaBoost obtained the lowest average rank of TPR in the borderline databases, but its average rank of TNR was much higher than the one of random forest.



(a) Safe-borderline (b) Borderline

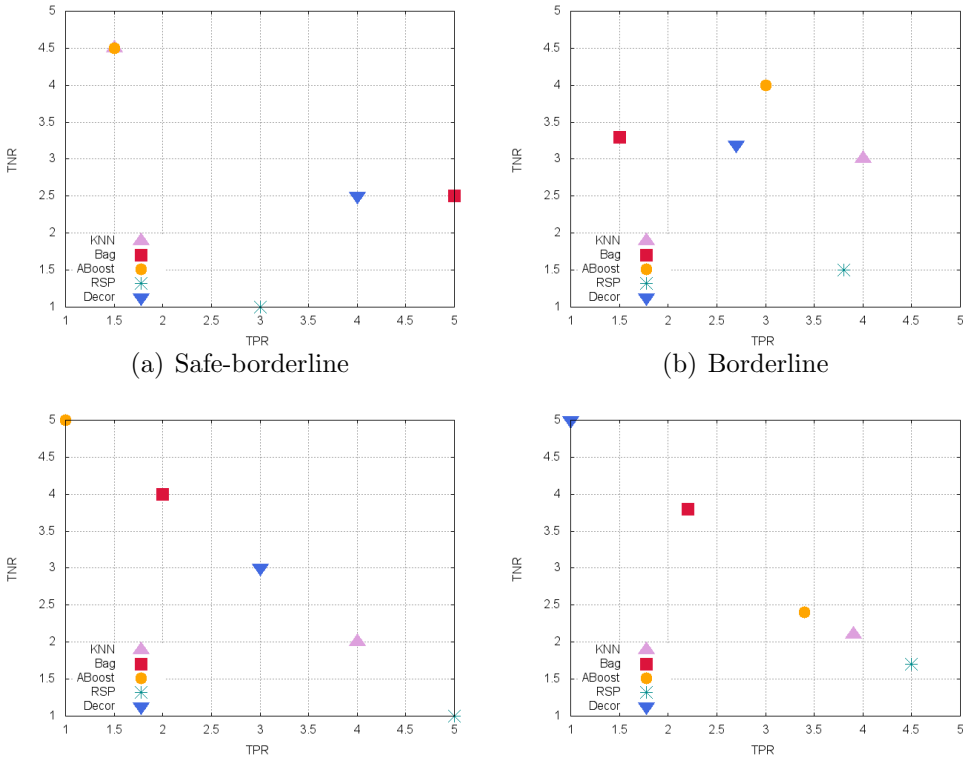
(c) Borderline-rare (d) Outlier

Figure 6: Average ranks of TPR vs. average ranks of TNR for the MLP-based ensembles

In the case of the outlier databases, Table 3 indicates that the best ensembles were BAGGING and AdaBoost, but the observation of results plotted in Figure 8 discloses that the latter performed much better on the positive class than the former. It is also worth pointing out that the random forest applied to the outlier data sets gave very poor performance results in terms of TPR, suggesting that this ensemble should not be used when a very large percentage of positive samples are outliers.

6. Concluding remarks and future work

This paper has addressed the problem of credit risk and bankruptcy prediction with classifier ensembles pursuing to investigate whether or not there exists any potential difference in their performance due to the distribution



(a) Safe-borderline (b) Borderline

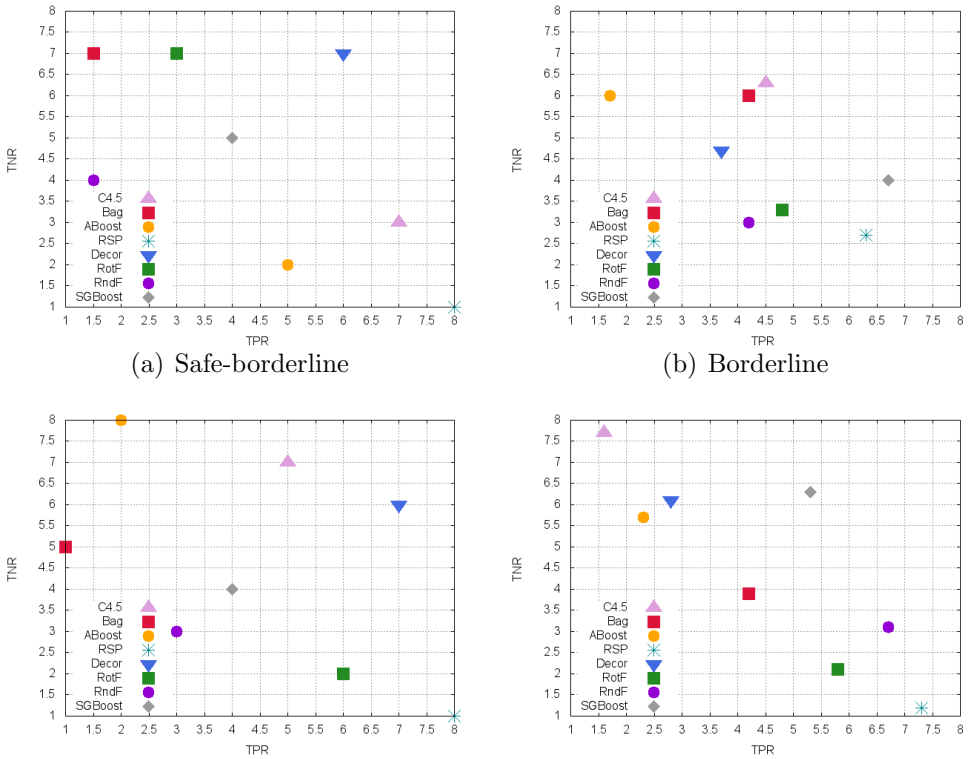
(c) Borderline-rare (d) Outlier

Figure 7: Average ranks of TPR vs. average ranks of TNR for the kNN-based ensembles

of sample types in a database. To this end, 14 real-life financial data sets have been characterized into five categories based on the prevalent type of samples in the positive class. Afterwards, a thorough pool of experiments has been carried out using seven well-established ensembles built with three base classifiers (the MLP with a hidden layer, the kNN decision rule and the C4.5 decision tree) commonly used in this application field.

The analysis on each category of databases has shown that the performance of any ensemble configuration indeed depends on the types of samples available in the data set. This finding can be especially useful when one has decide which classifier to apply for a particular problem in hand, thus avoiding to choose by a trial-and-error approach the most appropriate prediction model.

For future research, a natural extension to this work will consist in devel-



(a) Safe-borderline (b) Borderline

(c) Borderline-rare (d) Outlier

Figure 8: Average ranks of TPR vs. average ranks of TNR for the C4.5-based ensembles

oping a meta-learning framework that should be viewed as a decision support tool based on the characteristics of each database for the design of classification systems with the capability of achieving the highest performance. Another avenue for further research is to compare the performance of ensembles across the type of credit data sets (i.e., retail credit data versus corporate credit data) and investigate whether or not there exists any correlation with the performance of ensembles based on the sample types.

Acknowledgment

This work was supported by the Generalitat Valenciana [grant number PROMETEOII/2014/062].

Appendix A. Summary of ensembles applied to financial data

This appendix reports a summary of papers that have dealt with some intrinsic data characteristics by using various configurations of ensembles in the field of credit risk and corporate bankruptcy prediction. Table A.4 shows that most works have addressed the DistBI problems, whereas only a few have faced the FeatBI problems or a combination of both (DistBI+FeatBI).

Table A.4: Summary of ensemble models used on financial data to deal with intrinsic data characteristics (IDC)

Paper	Ensemble model	Base learner	IDC
[34]	Bag, RndF, WA[1] , HCES[2], KappaP[3], SDP[4], OO[5], kNORA[6]	ANN, DT[7], SVM	DistBI
[56]	SubBag[8], ABoost	kNN, SVM, DT	DistBI+FeatBI
[30]	SubBag, TrBag[9], TrABoost[1]0	SVM	DistBI
[36]	EBC[11], RndF, ABoost, GBoostDT[12], XGBoostDT[13], LogRS[14]	DT	DistBI
[49]	Bag	DT	FeatBI
[37]	Bag	DT	DistBI
[42]	Bag	SVM	DistBI
[33]	ABoost, Bag, RSP, Decor, RotF	CDT[15] , LogR[16], MLP, SVM, DT	DistBI
[50]	XGBoostDT, Bag, ABoost, RndF	ANN, DT	FeatBI
[53]	EFM-C[17]	ANN, SVM, RndF, DT, NB[18]	DistBI+FeatBI
[43]	EFM-C	ANN, SVM, RndF, DT, NB	DistBI
[46]	CSBoost[19], XGBoostDT, ABoost, RndF	DT	DistBI
[44]	CADF[20], GBoostDT, RndF	DT	DistBI
[45]	GMBoost[21], ABoost, CSBoost	SVM	DistBI
[51]	ABoost, Bag, RndF, Stacking	DT, MLP, SVM, NB	FeatBI
[38]	RndF, Bag,	LLogR[22]	DistBI
[39]	Bag	SVM	DistBI
[54]	Bag	ANN	DistBI+FeatBI
[52]	Bag, Boost	DT	FeatBI
[32]	GBoostDT, RandF	DT	DistBI
[35]	DCS[23], WRandF[24], IBRandF[25], kNORA	DT	DistBI
[47]	SMBoost[26]	RVM[27]	DistBI
[55]	Bag, RSP	DT	DistBI+FeatBI
[83]	Bag, ABoost	LogR, LDA, DT, MLP, kNN	DistBI
[40]	Bag	LogR, DT	DistBI
[48]	Bag, RandS	NB, kNN, LDA, DT, ANN	FeatBI
[41]	Bag, Bag+Stacking, RndF, XGBoost[28]	SVM, GPC[29], LogR	DistBI

[1]Weighted average for combining the base classifiers (WA)

[2]Hill climbing ensemble selection (HCES)

[3]Kappa prunning (KappaP)

[4]Semi-definite programming (SDP)

[5]Orientation ordering (OO)

[6]k nearest Oracle (kNORA)

[7]Decision tree (DT)
[8]Sub-BAGGING (SubBag)
[9]Transfer learning classifier ensemble with Bagging (TrBag)
[10]Transfer learning classifier ensemble with AdaBoost (TrABoost)
[11]Extended balance cascade (EBCA)
[12]Gradient boosting decision tree (GBoostDT)
[13]Boosted trees with extreme gradient boosting (XGBoostDT)
[14]Logistic regresion stacking (LogRS)
[15]Credal decision tree (CDT)
[16]Logistic regression (LogR)
[17]Ensemble fusion method with consensus (EFM-C)
[18]Naive Bayes (NB)
[19]Cost-senstitive boosting (CSBoost)
[20]Correlated-adjusted decision forest (CADF)
[21]Geometric mean boosting (GMBoost)
[22]Lasso-Logistic regression (LLogR)
[23]Dynamic classifier selection (DCS)
[24]Weighted random forests (WRandF)
[25]Improved balanced random forests (IBRandF)
[26]Soft margin boosting (SMBoost)
[27]Relevance vector machine (RVM)
[28]Extreme gradient boosting (XGBoost)
[29]Gaussian process classifier (GPC)

Appendix B. Full set of results

This appendix provides the results in terms of AUC, true-positive rate and true-negative rate achieved by each ensemble configuration over the databases included in the experiments. The first three columns in Tables B.5–B.7 are for the safe data sets, the following three are for the borderline ones, and the last column is for the safe-borderline database.The first seven columns in Tables B.8–B.10 are for the outlier data sets, and the last column is for the borderline-rare database.

References

[1] E. I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, J. Finan. 23 (1968) 589–609.

Table B.5: AUC for the safe, borderline and safe-borderline data sets

	Australian	Finland	SabiSPQ	Polish	German	Thomas	Japanese
MLP	0.923	0.963	0.863	0.804	0.780	0.622	0.924
Bag(MLP)	0.928	0.973	0.893	0.797	0.789	0.640	0.928
ABoost(MLP)	0.902	0.955	0.876	0.794	0.736	0.601	0.903
RSP(MLP)	0.924	0.973	0.905	0.816	0.784	0.633	0.924
Decor(MLP)	0.926	0.966	0.873	0.809	0.790	0.628	0.929
kNN	0.911	0.942	0.857	0.779	0.707	0.608	0.915
Bag(kNN)	0.888	0.936	0.881	0.825	0.690	0.616	0.872
ABoost(kNN)	0.911	0.942	0.857	0.779	0.707	0.583	0.913
RSP(kNN)	0.924	0.957	0.933	0.845	0.771	0.622	0.917
Decor(kNN)	0.899	0.957	0.870	0.836	0.719	0.619	0.905
C4.5	0.862	0.922	0.892	0.735	0.677	0.561	0.857
Bag(C4.5)	0.929	0.974	0.951	0.836	0.772	0.627	0.924
ABoost(C4.5)	0.910	0.978	0.944	0.822	0.748	0.616	0.915
RSP(C4.5)	0.921	0.971	0.948	0.829	0.762	0.583	0.915
Decor(C4.5)	0.917	0.975	0.942	0.844	0.758	0.587	0.908
RotF(C4.5)	0.932	0.978	0.951	0.852	0.776	0.602	0.920
RndF(C4.5)	0.936	0.982	0.956	0.858	0.794	0.637	0.936
SGBoost(C4.5)	0.933	0.972	0.938	0.814	0.782	0.635	0.921

Table B.6: True-positive rate for the safe, borderline and safe-borderline data sets

	Australian	Finland	SabiSPQ	Polish	German	Thomas	Japanese
MLP	0.857	0.868	0.756	0.673	0.557	0.000	0.892
Bag(MLP)	0.857	0.896	0.773	0.730	0.477	0.000	0.892
ABoost(MLP)	0.857	0.872	0.756	0.767	0.557	0.000	0.892
RSP(MLP)	0.785	0.896	0.763	0.749	0.160	0.000	0.879
Decor(MLP)	0.890	0.896	0.769	0.830	0.513	0.016	0.896
kNN	0.879	0.792	0.714	0.741	0.263	0.133	0.905
Bag(kNN)	0.785	0.800	0.790	0.768	0.467	0.334	0.784
ABoost(kNN)	0.879	0.792	0.714	0.741	0.263	0.356	0.905
RSP(kNN)	0.788	0.844	0.775	0.768	0.093	0.025	0.838
Decor(kNN)	0.805	0.820	0.773	0.759	0.437	0.248	0.814
C4.5	0.808	0.915	0.844	0.738	0.477	0.151	0.831
Bag(C4.5)	0.861	0.919	0.846	0.725	0.477	0.206	0.868
ABoost(C4.5)	0.824	0.932	0.858	0.759	0.513	0.201	0.838
RSP(C4.5)	0.799	0.922	0.825	0.749	0.284	0.003	0.811
Decor(C4.5)	0.827	0.932	0.852	0.750	0.441	0.158	0.834
RotF(C4.5)	0.857	0.930	0.835	0.748	0.459	0.086	0.865
RndF(C4.5)	0.874	0.937	0.850	0.748	0.419	0.223	0.868
SGBoost(C4.5)	0.859	0.912	0.826	0.727	0.436	0.073	0.861

[2] J. A. Ohlson, Financial ratios and the probabilistic prediction of bankruptcy, J. Account. Res. 18 (1980) 109–131.

[3] C. V. Zavgren, Assessing the vulnerability to failure of American in-

Table B.7: True-negative rate for the safe, borderline and safe-borderline data sets

	Australian	Finland	SabiSPQ	Polish	German	Thomas	Japanese
MLP	0.890	0.936	0.930	0.674	0.834	0.998	0.843
Bag(MLP)	0.867	0.932	0.966	0.687	0.881	1,000	0.863
ABoost(MLP)	0.890	0.928	0.930	0.680	0.834	0.998	0.843
RSP(MLP)	0.885	0.936	0.945	0.656	0.966	1,000	0.843
Decor(MLP)	0.867	0.940	0.928	0.601	0.881	0.998	0.860
kNN	0.836	0.932	0.833	0.758	0.933	0.972	0.821
Bag(kNN)	0.828	0.916	0.820	0.781	0.777	0.808	0.829
ABoost(kNN)	0.836	0.932	0.833	0.758	0.933	0.773	0.821
RSP(kNN)	0.906	0.932	0.968	0.766	0.983	0.994	0.896
Decor(kNN)	0.833	0.920	0.780	0.766	0.806	0.905	0.829
C4.5	0.872	0.940	0.906	0.698	0.847	0.940	0.874
Bag(C4.5)	0.874	0.952	0.946	0.784	0.867	0.912	0.857
ABoost(C4.5)	0.867	0.940	0.928	0.797	0.810	0.900	0.877
RSP(C4.5)	0.904	0.950	0.974	0.740	0.940	0.998	0.899
Decor(C4.5)	0.863	0.948	0.937	0.791	0.870	0.936	0.857
RotF(C4.5)	0.868	0.949	0.965	0.793	0.883	0.976	0.857
RndF(C4.5)	0.866	0.951	0.961	0.817	0.916	0.919	0.871
SGBoost(C4.5)	0.870	0.936	0.960	0.723	0.892	0.982	0.866

Table B.8: AUC for the outlier and borderline-rare data sets

	Polish-5th	Polish-4th	Iranian	Polish-3rd	Polish-2nd	Polish-1st	Taiwan
MLP	0.692	0.591	0.747	0.532	0.567	0.409	0.707
Bag(MLP)	0.764	0.677	0.787	0.647	0.519	0.554	0.713
ABoost(MLP)	0.637	0.567	0.710	0.535	0.569	0.404	0.685
RSP(MLP)	0.702	0.608	0.769	0.539	0.544	0.543	0.722
Decor(MLP)	0.687	0.592	0.753	0.536	0.537	0.557	0.705
kNN	0.724	0.665	0.776	0.582	0.535	0.611	0.745
Bag(kNN)	0.706	0.591	0.754	0.564	0.558	0.562	0.669
ABoost(kNN)	0.715	0.658	0.661	0.575	0.538	0.604	0.658
RSP(kNN)	0.820	0.728	0.781	0.701	0.652	0.711	0.756
Decor(kNN)	0.712	0.656	0.826			0.638	0.731
C4.5	0.827	0.781	0.573	0.765	0.803	0.796	0.658
Bag(C4.5)	0.936	0.901	0.770	0.900	0.879	0.923	0.748
ABoost(C4.5)	0.915	0.883	0.758	0.865	0.841	0.870	0.717
RSP(C4.5)	0.920	0.874	0.727	0.876	0.857	0.892	0.745
Decor(C4.5)	0.918	0.843	0.751	0.862	0.840	0.884	0.720
RotF(C4.5)	0.930	0.885	0.753	0.884	0.888	0.919	0.766
RndF(C4.5)	0.933	0.891	0.824	0.888	0.878	0.911	0.766
SGBoost(C4.5)	0.870	0.819	0.787	0.827	0.839	0.864	0.762

dustrial firms: a logistic analysis, J. Bus. Finan. Account. 12 (1985) 19–45.

[4] L. N. Allen, L. C. Rose, Financial survival analysis of defaulted debtors, J. Oper. Res. Soc. 57 (2006) 630–636.

Table B.9: True-positive rate for the outlier and borderline-rare data sets

	Polish-5th	Polish-4th	Iranian	Polish-3rd	Polish-2nd	Polish-1st	Taiwan
MLP	0.000	0.000	0.000	0.000	0.000	0.000	0.396
Bag(MLP)	0.000	0.000	0.000	0.000	0.000	0.000	0.387
ABoost(MLP)	0.000	0.000	0.000	0.000	0.000	0.000	0.396
RSP(MLP)	0.002	0.000	0.000	0.000	0.000	0.000	0.278
Decor(MLP)	0.002	0.000	0.000	0.000	0.000	0.000	0.377
kNN	0.034	0.002	0.000	0.000	0.000	0.022	0.318
Bag(kNN)	0.132	0.060	0.260	0.038	0.025	0.055	0.387
ABoost(kNN)	0.034	0.002	0.300	0.000	0.000	0.022	0.391
RSP(kNN)	0.024	0.000	0.020	0.000	0.000	0.011	0.264
Decor(kNN)	0.161	0.099	0.320			0.081	0.343
C4.5	0.602	0.422	0.124	0.403	0.465	0.515	0.362
Bag(C4.5)	0.588	0.379	0.096	0.346	0.404	0.502	0.385
ABoost(C4.5)	0.594	0.406	0.204	0.381	0.426	0.531	0.381
RSP(C4.5)	0.454	0.188	0.014	0.132	0.208	0.283	0.288
Decor(C4.5)	0.586	0.413	0.118	0.397	0.464	0.515	0.350
RotF(C4.5)	0.510	0.245	0.066	0.250	0.308	0.462	0.357
RndF(C4.5)	0.346	0.132	0.164	0.166	0.182	0.296	0.371
SGBoost(C4.5)	0.447	0.332	0.026	0.335	0.448	0.484	0.370

Table B.10: True-negative rate for the outlier and borderline-rare data sets

	Polish-5th	Polish-4th	Iranian	Polish-3rd	Polish-2nd	Polish-1st	Taiwan
MLP	1.000	1.000	1.000	1.000	1.000	1.000	0.936
Bag(MLP)	0.999	1.000	1.000	1.000	1.000	1.000	0.941
ABoost(MLP)	1.000	1.000	1.000	1.000	1.000	1.000	0.936
RSP(MLP)	1.000	1.000	1.000	1.000	1.000	1.000	0.960
Decor(MLP)	1.000	1.000	1.000	1.000	1.000	1.000	0.944
kNN	0.998	1.000	0.997	1.000	1.000	1.000	0.952
Bag(kNN)	0.979	0.981	0.968	0.983	0.987	0.989	0.843
ABoost(kNN)	0.998	1.000	0.965	1.000	1.000	1.000	0.826
RSP(kNN)	0.999	1.000	0.999	1.000	1.000	1.000	0.961
Decor(kNN)	0.953	0.953	0.964			0.970	0.935
C4.5	0.985	0.988	0.983	0.989	0.994	0.994	0.929
Bag(C4.5)	0.993	0.997	0.992	0.996	0.999	0.999	0.933
ABoost(C4.5)	0.990	0.993	0.983	0.994	0.997	0.997	0.898
RSP(C4.5)	0.995	1.000	0.997	1.000	1.000	1.000	0.960
Decor(C4.5)	0.990	0.997	0.991	0.990	0.993	0.996	0.931
RotF(C4.5)	0.994	0.999	0.994	0.999	1.000	1.000	0.953
RndF(C4.5)	0.994	0.997	0.995	0.998	0.999	0.999	0.943
SGBoost(C4.5)	0.987	0.992	0.992	0.992	0.994	0.995	0.940

[5] T.-S. Lee, C.-C. Chiu, Y.-C. Chou, C.-J. Lu, Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Comput. Stat. Data An. 50 (2006) 1113–1130.

[6] G. V. Karels, A. J. Prakash, Multivariate normality and forecasting of

business bankruptcy, J. Bus. Finan. Account. 14 (1987) 573–593.

[7] M. Mihalovic, Performance comparison of multiple discriminant analysis and logit models in bankruptcy prediction, Econ. Sociol. 9 (2016) 101–118.

[8] K.-S. Shin, T. S. Lee, H.-J. Kim, An application of support vector machines in bankruptcy prediction model, Expert Syst. Appl. 28 (2005) 127–135.

[9] Y. Ding, X. Song, Y. Zen, Forecasting financial condition of Chinese listed companies based on support vector machine, Expert Syst. Appl. 34 (2008) 3081–3089.

[10] B. E. Erdogan, Prediction of bankruptcy using support vector machines: an application to bank bankruptcy, J. Stat. Comput. Sim. 83 (2013) 1543–1555.

[11] M.-J. Kim, I. Han, The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms, Expert Syst. Appl. 25 (2003) 637–646.

[12] T. Lensberg, A. Eilifsen, T. E. McKee, Bankruptcy theory development and classification via genetic programming, Eur. J. Oper. Res. 169 (2006) 677–697.

[13] E. Acosta-González, F. Fernández-Rodríguez, Forecasting financial failure of firms via genetic algorithms, Comput. Econ. 43 (2014) 133–157.

[14] A. F. Atiya, Bankruptcy prediction for credit risk using neural networks: A survey and new results, IEEE T. Neural Networ. 12 (2001) 929–935.

[15] L. Sun, P. P. Shenoy, Using Bayesian networks for bankruptcy prediction: Some methodological issues, Eur. J. Oper. Res. 180 (2007) 738–753.

[16] A. Khashman, Credit risk evaluation using neural networks: Emotional versus conventional models, Appl. Soft Comput. 11 (2011) 5477–5484.

[17] L. Cleofas-Sánchez, V. García, A. I. Marqués, J. S. Sánchez, Financial distress prediction using the hybrid associative memory with translation, Appl. Soft Comput. 44 (2016) 144–152.

[18] D. Zhao, C. Huang, Y. Wei, F. Yu, M. Wang, H. Chen, An effective computational model for bankruptcy prediction using kernel extreme learning machine approach, Comput. Econ. (2016) 1–17.

[19] R. Slowinski, C. Zopounidis, Application of the rough set approach to evaluation of bankruptcy risk, Intell. Syst. Account. Financ. Manag. 4 (1995) 27–41.

[20] T. E. Mckee, Developing a bankruptcy prediction model via rough sets theory, Intell. Syst. Account. Financ. Manag. 9 (2000) 159–173.

[21] J. D. Cabedo, J. M. Tirado, Rough sets and discriminant analysis techniques for business default forecasting, Fuzzy Econ. Rev. 20 (2015) 3–37.

[22] D. Delen, C. Kuzey, A. Uyar, Measuring firm performance using financial ratios: A decision tree approach, Expert Syst. Appl. 40 (2013) 3970–3983.

[23] S. Y. Kim, A. Upneja, Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models, Econ. Model. 36 (2014) 354–362.

[24] D. West, S. Dellana, J. Qian, Neural network ensemble strategies for financial decision applications, Comput. Oper. Res. 32 (2005) 2543–2559.

[25] M. Doumpos, C. Zopounidis, Model combination for credit risk assessment: A stacked generalization approach, Ann. Oper. Res. 151 (2007) 289–306.

[26] E. Alfaro, N. García, M. Gámez, D. Elizondo, Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks, Decis. Support Syst. 45 (2008) 110–122.

[27] J. Sun, H. Li, Financial distress prediction using support vector machines: Ensemble vs. individual, Appl. Soft Comput. 12 (2012) 2254–2265.

[28] K. Pluto, D. Tasche, Estimating probabilities of default for low default portfolios, in: The Basel II Risk Parameters: Estimation, Validation, and Stress Testing, Springer, 2006, pp. 75–101.

[29] B. Twala, Combining classifiers for credit risk prediction, J. Sys. Scien. and Sys. Eng. 18 (2009) 292–311.

[30] J. Xiao, R. Wang, G. Teng, Y. Hu, A transfer learning based classifier ensemble model for customer credit scoring, in: Proc. IEEE Seventh International Joint Conference on Computational Sciences and Optimization, Beijing, China, 2014, pp. 64–68.

[31] S. Das, S. Datta, B. B. Chaudhuri, Handling data irregularities in classification: Foundations, trends, and future challenges, Pattern Recogn. 81 (2018) 674–693.

[32] I. Brown, C. Mues, An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Expert Syst. Appl. 39 (2012) 3446–3453.

[33] J. Abellán, J. G. Castellano, A comparative study on base classifiers in ensemble methods for credit scoring, Expert Syst. Appl. 73 (2017) 1–10.

[34] X. Feng, Z. Xiao, B. Zhong, J. Qiu, Y. Dong, Dynamic ensemble classification for credit scoring using soft probability, Appl. Soft Comput. 65 (2018) 139–151.

[35] J. Xiao, L. Xie, C. He, X. Jiang, Dynamic classifier ensemble model for customer classification with imbalanced class distribution, Expert Syst. Appl. 39 (2012) 3668–3675.

[36] H. He, W. Zhang, S. Zhang, A novel ensemble method for credit scoring: Adaption of different imbalance ratios, Expert Syst. Appl. 98 (2018) 105–117.

[37] J. Sun, J. Lang, H. Fujita, H. Li, Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates, Inform. Sci. 425 (2018) 76–91.

[38] H. Wang, Q. Xu, L. Zhou, Large unbalanced credit scoring using Lasso-Logistic regression ensemble, PLoS One 10 (2015) 1–20.

[39] J. Sun, Z. Shang, H. Li, Imbalance-oriented SVM methods for financial distress prediction: a comparative study among the new SB-SVM-ensemble method and traditional methods, J. Oper. Res. Soc. 65 (2014) 1905–1919.

[40] F. Louzada, O. Anacleto-Junior, C. Candolo, J. Mazucheli, Poly-bagging predictors for classification modelling for credit scoring, Expert Syst. Appl. 38 (2011) 12717–12720.

[41] Y. Xia, C. Liu, B. Da, F. Xie, A novel heterogeneous ensemble credit scoring model based on bstacking approach, Expert Syst. Appl. 93 (2018) 182–199.

[42] L. Yu, R. Zhou, L. Tang, R. Chen, A dbn-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data, Appl. Soft Comput. 69 (2018) 192–202.

[43] M. Ala'raj, M. F. Abbod, Classifiers consensus system approach for credit scoring, Knowl.-Based Syst. 104 (2016) 89–105.

[44] R. Florez-Lopez, J. M. Ramon-Jeronimo, Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. a correlated-adjusted decision forest proposal, Expert Syst. Appl. 42 (2015) 5737–5753.

[45] M.-J. Kim, D.-K. Kang, H. B. Kim, Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, Expert Syst. Appl. 42 (2015) 1074–1082.

[46] M. Ziba, S. K. Tomczak, J. M. Tomczak, Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction, Expert Syst. Appl. 58 (2016) 93–101.

[47] S. Li, I. W. Tsang, N. S. Chaudhari, Relevance vector machine based infinite decision agent ensemble learning for credit risk analysis, Expert Syst. Appl. 39 (2012) 4947–4953.

[48] B. Twala, Multiple classifier application to credit risk assessment, Expert Syst. Appl. 37 (2010) 3326–3336.

[49] M. A. Muslim, A. Nurzahputra, B. Prasetiyo, Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction, in: International Conference on Information and Communications Technology, pp. 141–145.

[50] Y. Xia, C. Liu, Y. Li, N. Liu, A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, Expert Syst. Appl. 78 (2017) 225–241.

[51] F. N. Koutanaei, H. Sajedi, M. Khanbabaei, A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring, J. Retailing Cons. Ser. 27 (2015) 11–23.

[52] G. Wang, J. Ma, S. Yang, An improved boosting based on feature selection for corporate bankruptcy prediction, Expert Syst. Appl. 41 (2014) 2353–2361.

[53] M. Ala'raj, M. F. Abbod, A new hybrid ensemble credit scoring model based on classifiers consensus system approach, Expert Syst. Appl. 64 (2016) 36–55.

[54] J.-J. Liao, C.-H. Shih, T.-F. Chen, M.-F. Hsu, An ensemble-based model for two-class imbalanced financial problem, Econ. Model. 37 (2014) 175–183.

[55] G. Wang, J. Ma, L. Huang, K. Xu, Two credit scoring models based on dual strategy ensemble trees, Knowl.-Based Syst. 26 (2012) 61–68.

[56] G. Paleologo, A. Elisseeff, G. Antonini, Subagging for credit scoring models, Eur. J. Oper. Res. 201 (2010) 490–499.

[57] K. Napierala, J. Stefanowski, S. Wilk, Learning from imbalanced data in presence of noisy and borderline examples, in: Proc. 7th International Conference on Rough Sets and Current Trends in Computing, Warsaw, Poland, 2010, pp. 158–167.

[58] K. Napierala, J. Stefanowski, The influence of minority class distribution on learning from imbalance data, in: Proc. 7th International Conference on Hybrid Artificial Intelligence Systems, Salamanca, Spain, 2012, pp. 139–150.

[59] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: One-sided selection, in: Proc. 14th International Conference on Machine Learning, Nashville, TN, 1997, pp. 179–186.

[60] K. Napierala, J. Stefanowski, Types of minority class examples and their influence on learning classifiers from imbalanced data, J. Intell. Inf. Syst. 46 (2016) 563–597.

[61] J. A. Sáez, B. Krawczyk, M. Woźniak, Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets, Pattern Recogn. 57 (2016) 164–178.

[62] B. Krawczyk, M. Woniak, F. Herrera, Weighted one-class classification for different types of minority class examples in imbalanced data, in: Proc. IEEE Symposium on Computational Intelligence and Data Mining, Piscataway, NJ, 2014, pp. 337–344.

[63] P. Skryjomski, B. Krawczyk, Influence of minority class instance types on SMOTE imbalanced data oversampling, in: Proc. 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications, volume 74, Skopje, Macedonia, 2017, pp. 7–21.

[64] M. Lichman, UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml`, 2013. School of Information and Computer Sciences, University of California, Irvine, CA.

[65] P. Du Jardin, Financial failure forecasting and neural networks: the contribution of variable selection techniques, Ph.D. thesis, Edhec Business School, Université Nice Sophia Antipolis, France, 2007. `https://tel.archives-ouvertes.fr/tel-00475200`.

[66] W. Pietruszkiewicz, Dynamical systems and nonlinear Kalman filtering applied in classification, in: Proc. 7th IEEE International Conference on Cybernetic Intelligent Systems, London, UK, 2008, pp. 263–268.

[67] L. C. Thomas, D. B. Edelman, J. N. Crook, Credit Scoring and Its Applications, SIAM, Philadelphia, PA, 2002.

[68] I.-C. Yeh, C. hui Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Syst. Appl. 36 (2009) 2473–2480.

[69] H. Sabzevari, M. Soleymani, E. Noorbakhsh, A comparison between statistical and data mining methods for credit scoring in case of limited

available data, in: Proc. 3rd CRC Credit Scoring Conference, Edinburgh, UK, 2007.

[70] V. García, A. I. Marqués, J. S. Sánchez, An insight into the experimental design for credit risk and corporate bankruptcy prediction systems, J. Intell. Inf. Syst. 44 (2015) 159–189.

[71] J. Caouette, E. Altman, P. Narayanan, R. Nimmo, Managing Credit Risk: The Great Challenge for Global Financial Markets, Wiley, Hoboken, NJ, 2008.

[72] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.

[73] Y. Freund, R. E. Schapire, Experiments with a new boosting algorithm, in: Proc. 13th International Conference on Machine Learning, Bari, Italy, 1996, pp. 148–156.

[74] J. H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (2002) 367–378.

[75] T. K. Ho, The random subspace method for constructing decision forests, IEEE T. Pattern Anal. Mach. Intell. 20 (1998) 832–844.

[76] J. J. Rodríguez, L. I. Kuncheva, C. J. Alonso, Rotation forest: A new classifier ensemble method, IEEE T. Pattern Anal. Mach. Intell. 28 (2006) 1619–1630.

[77] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[78] P. Melville, R. J. Mooney, Creating diversity in ensembles using artificial data, Inform. Fusion 6 (2005) 99–111.

[79] J. Stefanowski, Dealing with data difficulty factors while learning from imbalanced data, in: Challenges in Computational Statistics and Data Mining, Springer, 2016, pp. 333–363.

[80] C. Drummond, R. C. Holte, Cost curves: An improved method for visualizing classifier performance, Mach. Learn. 65 (2006) 95–130.

[81] J. M. Lobo, J. Alberto, R. Raimundo, AUC: a misleading measure of the performance of predictive distribution models, Global Ecol. Biogeogr. 17 (2008) 145–151.

[82] D. J. Hand, Measuring classifier performance: a coherent alternative to the area under the ROC curve, Mach. Learn. 77 (2009) 103–123.

[83] S. Finlay, Multiple classifier architectures and their application to credit risk assessment, Eur. J. Oper. Res. 210 (2011) 368–378.