Noname manuscript No.
(will be inserted by the editor)

An insight into the experimental design for credit risk and corporate bankruptcy prediction systems

Vicente García · Ana I. Marqués · J. Salvador Sánchez

Received: date / Accepted: date

Abstract Over the last years, it has been observed an increasing interest of the finance and business communities in any application tool related to the prediction of credit and bankruptcy risk, probably due to the need of more robust decision-making systems capable of managing and analyzing complex data. As a result, plentiful techniques have been developed with the aim of producing accurate prediction models that are able to tackle these issues. However, the design of experiments to assess and compare these models has attracted little attention so far, even though it plays an important role in validating and supporting the theoretical evidence of performance. The experimental design should be done carefully for the results to hold significance; otherwise, it might be a potential source of misleading and contradictory conclusions about the benefits of using a particular prediction system. In this work, we review more than 140 papers published in refereed journals within the period 2000–2013, putting the emphasis on the bases of the experimental design in credit scoring and bankruptcy prediction applications. We provide some caveats and guidelines for the usage of databases, data splitting methods, performance evaluation metrics and hypothesis testing procedures in order to converge on a systematic, consistent validation standard.

Keywords Credit risk · Corporate bankruptcy · Experimental design · Data splitting · Performance metric · Statistical test

V. García
Department of Electrical Engineering and Computing, Universidad Autónoma de Ciudad Juárez, 32310 Ciudad Juárez, Chihuahua (Mexico)
E-mail: vicente.jimenez@uacj.mx

A.I. Marqués
Department of Business Administration and Marketing, Universitat Jaume I, 12071 Castelló de la Plana (Spain)
E-mail: imarques@uji.es

J.S. Sánchez
Institute of New Imaging Technologies, Department of Computer Languages and Systems, Universitat Jaume I, 12071 Castelló de la Plana (Spain)
Tel.: +34-964-728350
Fax: +34-964-728435
E-mail: sanchez@uji.es

1 Introduction

Credit risk and corporate bankruptcy prediction constitutes an application domain of major interest for banks and financial institutions because erroneous decisions may lead to very important costs (Horcher 2005). This is the reason why the development of a great variety of strategies to implement reliable prediction models has attracted considerable attention both from academicians and financial analysts over the last decades. These range from very traditional statistical techniques (e.g., weight of evidence, logistic regression, discriminant analysis, multivariate adaptive regression splines, probit analysis) to more sophisticated computational intelligence paradigms (e.g., neural networks, support vector machines, evolutionary computing, fuzzy algorithms, expert systems) and operations research methodologies (e.g., mathematical programming, multi-criteria decision making methods).

Prediction of credit risk and bankruptcy can be performed through the generation of models, which are usually based on a binary classification approach, in order to distinguish potential defaulters (bankrupters) from non-defaulters (non-bankrupters). From a practical point of view, classification refers to the assignment of a finite set of samples to predefined classes based on a number of observed variables or attributes (Thomas et al 2002). For instance, the input of a credit scoring system may consist of a collection of historical information that describes socio-demographic characteristics and economic conditions of the applicant, and the classification model produces the output in terms of the customer creditworthiness.

Despite the growing interest in developing more accurate prediction models, the issue of how these models should be evaluated and their results thoroughly validated has not been investigated sufficiently so far. An example of this paradox is the considerable number of surveys that summarize the many techniques proposed in the literature and/or compare their performance results, but they do not concentrate on how the experiments have been designed. Just to cite a few recent examples, Crook et al (2007) review a selection of statistical models, mathematical programming and soft computing techniques for consumer credit risk assessment. Ravi Kumar and Ravi (2007) present an extensive analysis of statistical and intelligent methods applied to the prediction of corporate bankruptcy risk in the period 1968–2005, highlighting the source of data, financial ratios and country of origin. Verikas et al (2010) focus their review on how to combine different soft computing techniques to derive hybrid and ensemble-based bankruptcy prediction models. Lin et al (2012) provide a statistical survey of machine learning papers published between 1995 and 2010 in the realm of credit scoring and bankruptcy prediction, summing up the data sets and comparing the performance of several methods with baseline classifiers. Abdou and Pointon (2011) present a literature review of works related to credit scoring applications in various areas, with the aim of investigating how this field has grown in importance over the last decades and also identifying the primary factors in the construction of a credit scoring model. Sadatrasoul et al (2013) give a comprehensive review of studies where data mining techniques have been applied to credit scoring from 2000 to 2012.

Unfortunately, none of these surveys provides a deep insight into the process of experimentation and validation, even though it is widely accepted that the proper design of experiments constitutes a paramount factor to ensure a complete understanding and testing of the performance of the prediction models developed (Cohen 1995). At least four key components should be defined carefully in order to draw well-founded conclusions from the results: the experimental data, the data splitting methods, the performance evaluation metrics and the statistical tests of significance. Nevertheless, the configuration of these ele-

ments is often done in a blind manner within the experimental framework for the prediction of credit risk and bankruptcy.

This application area presents certain dominant characteristics that make the design of experiments especially critical and challenging, with a number of particularities that differ from other real-life applications in two aspects. First some problems are recurrent, and second they appear in combination with other complexities. The following list reports some of the most significant features of credit risk and corporate bankruptcy prediction.

- Data sets are typically characterized by highly imbalanced class distribution with a scarcity of default observations, which is often referred to as the low-default portfolio problem.
- Most data sets range from small to medium in size (in terms of the number of examples).
- In general, the costs of false negative errors and false positive errors are asymmetric.
- Data sets often include records with missing values.
- Multiple conflicting evaluation criteria have to be usually taken into account.
- Samples are described by both quantitative and qualitative attributes (independent variables), with some of them being irrelevant and/or redundant.
- It is frequent enough to find noisy, atypical examples in the data sets.

Some of these characteristics should carefully be taken into consideration when designing the experiments because there is evidence that they may affect the experimental results strongly. For instance, the imbalanced nature of data in a credit risk application and the asymmetric misclassification costs require the use of performance evaluation metrics that are not biased towards the majority class. Also, the multiple (usually conflicting) criteria may give rise to contradictory predictions if an inappropriate single measure is used to evaluate the models. On the other hand, the size of data sets determines how to split the data, and this becomes even more important with a high imbalance ratio. In our opinion, it seems clear enough the importance of keeping in mind the particularities of this application area in order to define a comprehensive experimental methodology.

Accordingly, this work conducts a systematic review of more than 140 papers published in refereed journals within the period between 2000 and 2013. The purpose of this survey is studying how the experiments have been designed and the results validated in the field of credit risk and corporate bankruptcy prediction. To this end, each of the four aforementioned experimental components will be analyzed, while discussing the limitations of the standard configurations used in current practice and providing suggestions to establish a more robust experimental methodology that can help authors enhance their studies. However, we would like to elucidate that our analysis does not intend to criticize any previous research efforts.

Henceforth, this paper is organized as follows. Section 2 gives an overview of the research methodology we have adopted to conduct our investigation. Section 3 analyzes the experimental databases in terms of their sources and sizes. Section 4 discusses the data splitting methods and suggests how they should be applied in order to yield consistent results. Section 5 outlines the criteria used to measure the model performance and points out the adequacy of each one depending on the characteristics of the databases. The most common statistical methods used to test the significance of performance results are studied in Section 6. Two simple experimental scenarios are included in Section 7 in order to stress the different performance results and the conclusions that can be drawn from them depending on the experimental methodology adopted. Next, Section 8 proffers a set of caveats and simple guidelines for better experimental design and validation of credit risk and corporate bankruptcy prediction models. Finally, Section 9 remarks the main findings of our research.

2 Research methodology

Figure 1 illustrates an overall picture of the main steps involved in the research process followed for conducting our study. This process is based upon the suggestions given by Staples and Niazi (2007) and comprises two basic phases (each one with a sequence of steps): the *definition stage* to establish the purpose and the protocol for the research, and the *development stage* for collecting related papers, handling the relevant data, analyzing the results and drawing conclusions. It is worth noting that this process is not linear, rather it requires iteration, feedback and refinement.

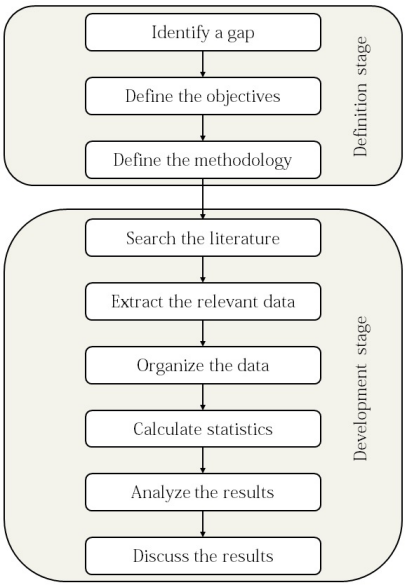

Fig. 1 Block diagram of the research process followed in this study

The definition stage consists of three steps. In the first one, we have to identify the need for a new research to cover a gap in the domain of study. Afterwards, the definition of the general objectives allows to better circumscribe the specific research to be undertaken, whereas the definition of the methodology aims at giving a formal and detailed protocol for the execution of the research. Both the identification of a gap and the definition of the objectives (the first two steps in Figure 1) have already been addressed in the previous section, while the definition of the research methodology will be discussed next in this section.

The first step of the development stage employs a particular search strategy to retrieve an initial list of publications that may be relevant to the objectives. Nonetheless, this process needs further refinement in order to exclude some papers that do not fulfill the research requirements completely and include some others that may be of interest to our study. In particular, the present investigation on experimental design for credit scoring and corporate bankruptcy risk prediction was carried out by cross-searching for related journal papers with the support of eight comprehensive bibliographic databases: ISI Web of Science, Google Scholar, IEEE Xplore, SpringerLink, Scopus, Inspec, ScienceDirect, and ACM Digital Library. Conference papers were excluded from the initial list of studies because in general,

the empirical work in this kind of publications appears to be much less exhaustive than in journal articles due to the lack of space; therefore, their inclusion could give rise to erroneous conclusions. Besides, the reference section in each of the retrieved papers was also scanned to add up some other relevant studies not included in the initial list of publications.

From the final list of 142 papers, the data extraction step was designed to collect data pertinent to the present study. For each article, we recorded the journal title and the year of publication, along with the databases, the data splitting procedures, the performance evaluation metrics and the hypothesis testing methods used in the experiments. Then all this information was organized in the form of a table to make easier the computation of statistics and the analysis of results.

We collected papers on credit risk and bankruptcy prediction from more than 50 scientific journals, which are mostly related to the fields of management, operations research, information and computer science, economics, and finance. Table 1 summarizes the journals with at least four articles included in the present study, reporting the number of papers along with the proportions and cumulative proportions for each journal. As can be observed, eight journals contribute with almost 60% of the total amount of papers in review, but one should not overlook the remaining publications because some relevant results might be missed out.

Table 1 Distribution of papers across journals.

Journal	#Papers	Proportion	Cumulative proportion
Expert Systems with Applications	44	30.99	30.99
European Journal of Operational Research	12	8.45	39.44
Journal of the Operational Research Society	6	4.23	43.67
Computers & Operations Research	4	2.82	46.49
Decision Support Systems	4	2.82	49.31
Knowledge-Based Systems	4	2.82	52.13
Information Sciences	4	2.82	54.95
Applied Soft Computing	4	2.82	57.77

3 Databases

The first component that has to be chosen carefully in the experimental design is the data with which to perform the experiments. As soon as one starts to set up the experimental protocol, several questions regarding the number of databases to use, the data set size, or the type of variables arise. Therefore, one should take care of all these questions in order to define an appropriate configuration of the experiments with the aim of maximizing the significance of the results.

From the literature review carried out, we have mainly observed two significant trends regarding the data used for the experiments. First, several works have employed benchmarking databases such as the extremely well-known Australian and German data sets, which can be taken from the UCI Machine Learning Database Repository (Bache and Lichman 2013). Even though these data sets are among the most widely used for credit scoring and bankruptcy prediction, many other studies have experimented with private databases collected by several local financial institutions, which are generally thought to face a specific application problem.

Each of these two options has its pros and cons. In the case of using benchmarking databases, the main advantage is that they allow future experimenters to make extensive

comparisons between different prediction models; however, these data sets may not be representative enough of the current socio-economic conditions and hence the experiments may lead to outdated and worthless conclusions. Conversely, application-oriented databases are mainly thought to tackle some particular real-world problems, but there may be difficulties to employ them for further comparisons. Also, it is worth stressing that many studies with private data do not include a complete description of the variables that comprise the samples, and even others do not provide the database size (e.g. Pavlenko and Chernyak 2010), the number of variables (e.g. Pavlenko and Chernyak 2010; Ben-David and Frank 2009), or the proportion of samples that belong to each class of the data set (e.g. Galindo and Tamayo 2000; Hoffmann et al 2002), thus making difficult to understand in depth the merits (or faults) and procedural issues of each model.

Because of the shortcomings related to the individual usage of either public or private data, it will be generally better to employ a mixture of both benchmarking and application-oriented databases. Nevertheless, as can be seen in Figure 2(a), only 13% of the papers reviewed have involved both types of databases in their experiments, while the rest is distributed between those that have employed only public databases and mainly those that have experimented with only private data sets. As can be seen, nearly a quarter of the studies have focused only on benchmarking databases, whereas about two thirds have used only data gathered from their own sources.

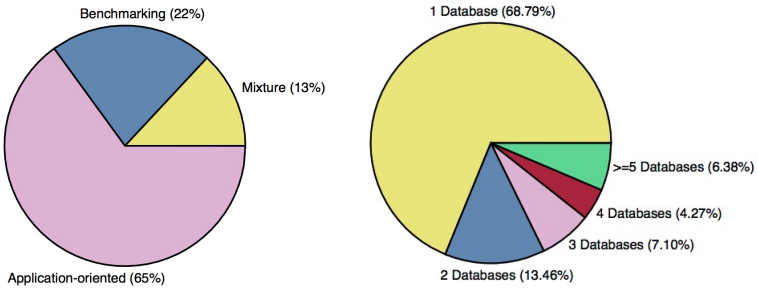

Fig. 2 Percentages of papers (a) using benchmarking, application-oriented or a mixture of databases, (b) as a function of the number of databases used in the experiments.

As a final comment, it should be remarked that more than 68% of the papers analyzed in this survey have limited their experiments to a single database (see Figure 2(b)); thus it is not possible to extrapolate any conclusions about the strengths and weaknesses of a prediction method to other data from different financial institutions. There exist supporting empirical evidences that it is preferable to use several different data sets for model evaluation rather than a unique database in order to draw significant and meaningful conclusions, but only 6.38% of the papers have included five or more data sets in their experiments.

3.1 Data set size

An important characteristic of the databases that should be analyzed in depth refers to their size, which is determined by both the number of examples and the number of attributes or

independent variables. A drawback common to most of the data sets used in these papers relates to the small sample size, which may produce a relatively high variance of any statistic calculated from them. In order to better understand how are the databases, we have classified them into three categories based on the number of samples available: small size (less than 1,000 samples), medium size (1,000 – 10,000 samples) and large size (more than 10,000 samples). The papers reviewed have considered more than 110 different databases, with 59 being small, 40 medium and 13 large sized. Table 8 in Appendix A provides a brief description of the databases sorted in ascending order by the total number of samples (N), and also reports the number of attributes (D) and the references to the papers that have conducted experiments over each database. In a significant number of studies, it is possible to observe that most data sets consist of a very small number of examples ($N < 200$), such as the case of the 160 electronics companies listed by the Taiwan Stock Exchange Corporation (Chen 2013), the Spanish non-life insurance database with 72 firms (Salcedo-Sanz et al 2005), the 118 bankrupt and non-bankrupt examples of Greek industries (Tsakonas et al 2006), the database of Jordanian commercial banks with 140 personal loan applications (Eletter et al 2010), or the financial data collected from 105 small companies in Romania (Cimpoeru 2011).

Figure 3 shows the number of papers per year as a function of the size of the data sets used in the experiments. It has to be noted that the Japanese, Australian and German databases have not been considered for this analysis because they have extensively been employed in many works and therefore their inclusion could distort any conclusions. As can be observed, there is a constant trend toward the use of small and medium sized databases across the period of study. However, it seems that experimentation with larger data sets has increased moderately over the last years. For instance, only three papers employed large databases from 2000 to 2008 (one in 2003, one in 2005 and one in 2007), but four articles published in 2013 have already used large sized data sets.

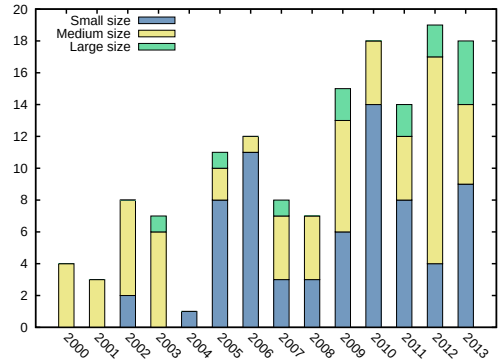

Fig. 3 Year-wise distribution of papers according to the data set size (N).

On the other hand, the values of N and D in Table 8 suggest that there does not exist a strong correlation between the number of variables and the sample size of the databases used in the reviewed papers. As a way of checking our claim, we have computed the Pearson's correlation coefficient between these two features, which certainly corroborates a low degree of correlation ($r = 0.27$). In fact, the number of attributes is in the range of 5 to 30 for 89% of small, 82% of medium and 46% of large sized databases. While this number of variables (5–

30) can be suitable for databases with more than 1000 samples, it can become a hindrance for databases with a limited number of samples because the performance of a prediction model decreases as the dimensionality increases. Therefore, it is important to examine the ratio of the number of samples to the number of variables because this can lead to problematic situations due to the so-called Hughes phenomenon (Hughes 1968), which states that the ratio of the number of samples to the number of attributes must be maintained at or above some minimum value to achieve accurate predictions. Although there is no strict guideline about what a sufficient data size is, Nagy (2004) claims that it should be around $10 \times D \times C$, where C is the number of classes in a problem. Unfortunately, several databases included in Table 8 do not fulfill this rule, such as the case of the Lithuanian database with 60 variables and 100 samples (Boguslauskas and Mileris 2009; Mileris 2010) or the Shanghai/Shenzhen Stock Exchange data with 30 variables and 153 samples (Li and Sun 2009).

4 Data splitting methods

The fundamental idea behind data splitting (or resampling) is very simple: we isolate one part of the data, learn on the rest, and then evaluate the model on the portion that was isolated. Briefly, data splitting methods are based on some form of partitioning of the available data into a training set for building the classifier or prediction model and a test set that will be used only for model assessment. In general, the larger the training set, the better the classifier; but also the larger the test set, the more accurate the performance estimate. In the case of credit risk and bankruptcy prediction, where the amount of data is usually very scarce, the resampling strategies become of great relevance for reliable model evaluation. Thus correctness of experimental results strongly depends on the selection of an appropriate resampling method, which in turn should be based on the data available for the experiments. As data are limited, one has to find a trade-off between the size of the training set and the size of the test set.

The data splitting procedures (Alpaydin 2010) that have mostly adopted by the papers analyzed in this work are:

- Holdout. The data set S is randomly split into two disjoint subsets, S_{tra} and S_{tst}; the model is built using the samples in S_{tra} and assessed on samples in S_{tst}.
- K-fold cross validation (CV). The data set S with N samples is randomly divided into K mutually exclusive subsets of approximately equal size, S_i, $i = 1, 2, \ldots, K$ ($K \leq N$). Each subset is in turn left out during model building; the model is trained on the union of the remaining $K-1$ subsets and predictions are obtained for the left out subset. After the K rounds of training and testing are complete, all the test set predictions are used to estimate the model performance.
- Leave-one-out (LOO). This is a particular case of K-fold cross validation with $K = N$, that is, a single sample is left out each time; at each round, $K-1$ cases are used for training and only one for testing.
- $K_1 \times K_2$-fold CV. The K_2-fold cross validation method is repeated K_1 times and then the model performance is obtained as the average of the $K_1 \times K_2$ estimates.

An important issue closely related to resampling is stratification, which ensures that the class distribution of the original data set is preserved in the training and test sets, that is, the prior class probabilities should be kept in all partitions. This avoids the potential problem of generating some subsets with no examples from one of the two classes (Forman and Scholz 2010). On the other hand, it has been demonstrated that stratification helps to reduce

the variance of the estimated performance (Kohavi 1995), especially for data sets with many classes (Sechidis et al 2011). Despite its relevance, a great majority of papers do not indicate whether or not they have used a stratified data splitting technique.

In the holdout method, the key question is how many samples should be left out for the test set. It has been observed that the holdout estimator tends to be too pessimistic because only a proportion of the data is used to build the model (Bischl et al 2012). Correspondingly, a variation of the holdout method, which partially alleviates this biased behavior, consists of replicating the partition into training and test sets several times in different random ways; the classifier is trained and tested for each partition and the performances averaged to yield an overall estimate, which is generally more reliable.

For K-fold cross validation, the question is how many subsets should be used. With a large number of subsets, the estimator will be very accurate, but the variance will be large. Conversely, with a reduced number of subsets, the variance will be small, but the estimator will be largely biased (i.e, too conservative) (Bischl et al 2012). Although $K = 5$ and $K = 10$ are common choices that perform reasonably well for data sets of different sizes, it is worth noting that for very small data sets, a bigger value of K (or even the leave-one-out method) may become slightly preferable in order to train on as many examples as possible.

Table 2 Distribution of papers across each data splitting strategy.

	Proportion	Cumulative proportion	Typical settings
Holdout	35.46	35.46	80/20, 70/30
K-fold CV	30.50	65.96	$K = 5, 10$
Repeated holdout	14.89	80.85	5, 10, 50, 100 times
LOO	7.04	87.89	–
$K_1 \times K_2$-fold CV	4.26	92,15	$5 \times 10, 10 \times 5$
Not specified	7.85	100	–

Table 2 provides the distribution of papers according to the usage of the most typical resampling procedures. A simple glance at this table reveals that single holdout and K-fold cross validation are indeed the most popular resampling algorithms in the field of credit risk and corporate bankruptcy prediction, being applied on nearly 66% of the articles. The repeated holdout method has been chosen in less than 15% of the papers, showing that not many researchers are aware of the need for multiple runs. Paradoxically, despite the small size of many of these databases, the leave-one-out estimator has been employed only in 7% of the studies. It has also to be noted that about 8% of the papers have not indicated the data splitting procedure used in their experiments, which makes quite difficult to figure out the correctness of the results and the consistency of the conclusions.

Another question that deserves to be analyzed is whether there exists any relationship between the data set size and the resampling method used. To this end, Table 3 reports how many articles have used a given data splitting method with small, medium and large databases. For instance, three different papers have employed the $K_1 \times K_2$-fold cross validation over small and medium sized data sets. Despite holdout and K-fold cross validation correspond to the resampling strategies with the lowest cost, they are the most widely-used methods even for small and medium sized databases. As can be observed, leave-one-out is applied when the data size is small because its computational burden is likely to be too high for databases with more than 1000 samples. Although the reduced number of papers that have experimented with large databases does not allow to draw any conclusions for this

category, it seems that the use of leave-one-out and $K_1 \times K_2$-fold cross validation has been discarded because of their high time-consuming nature.

Table 3 Number of papers per data splitting strategy and data size.

	Small	Medium	Large
Holdout	41	36	3
K-fold CV	48	44	3
Repeated holdout	10	13	5
LOO	9	1	0
$K_1 \times K_2$-fold CV	3	3	0

Even though the seeming relationship between data set size and resampling strategy, a more in-depth analysis of the papers shows that different authors have used different data splitting methods over a same database. This is especially obvious in the case of the Japanese, Australian and German databases, where holdout, K-fold cross validation and repeated holdout have all been applied equally. But this can also be found in many other data sets, such as the US bank database where Li et al (2008) applied the holdout method, Peng et al (2008, 2011) used 10-fold cross validation and Zhou et al (2011) employed the repeated holdout approach. In addition, some articles with experiments over various databases apply the same data splitting method regardless of the data size. For instance, Brown and Mues (2012) use holdout on five data sets with different sizes ranging from 547 to 7190 samples, and García et al (2012) apply 5-fold cross validation on eight data sets with sizes ranging from 240 to 5000 samples. This suggests that the choice of a particular resampling strategy is not always based on the size of data, but it may depend on the preferences of each author.

5 Performance evaluation metrics

The third component to be considered in the design of experiments involves how to assess the performance of the models tested on the data that have previously been picked out, that is, one has to select the performance evaluation measure (or a collection of them) that better fits the specific problem under consideration. In the framework of classification, the purpose of most performance evaluation metrics is to estimate how well the learned model predicts the correct class of new input samples, but not all of them are addressed to measure the same things. Therefore, the key question is to choose the most appropriate criteria that satisfy the special requirements for the problem in hand; otherwise, the results could lead to distorted conclusions since different metrics may yield different orderings of model performance (Hand 2012; Raeder et al 2012). In this section we examine the most popular scalar metrics used in the credit scoring and bankruptcy literature, restricting the discussion to the two-class problem because this is the most general case when undertaking these financial applications. For consistency with the common terminology used in this context, we will refer to the 'good' risk class (i.e., non-default, non-bankrupt) as positive and the 'bad' class (i.e., default, bankrupt) as negative.

Classification accuracy (acc) and its counterpart, the error rate, have been by far the most frequently employed indicators of performance in the papers reviewed (more than 88% of the papers include the accuracy or the error in their experiments). For a two-class problem, both these metrics can be derived from a 2×2 confusion or co-occurrence matrix as the one in Table 4, where columns represent the predicted class and rows indicate the true class;

Table 4 Confusion matrix for a two-class problem.

	Positive prediction	Negative prediction
Positive class	True Positives (TP)	False Negatives (FN)
Negative class	False Positives (FP)	True Negatives (TN)

each entry (i, j) contains the number of correct/incorrect predictions. Its diagonal contains the number of cases that have correctly been predicted for each class, while the off-diagonal elements indicate the number of samples that have been classified wrongly.

Both accuracy and error rate assume symmetric misclassification costs for the positive and negative classes (good observations being predicted as bad, and vice versa). This is the reason why approximately 41% of the papers also measure the error on each individual class by using the so-called type-I and type-II errors. Type-I error (or miss) is the rate of bad cases being categorized as good; when this happens, the misclassified bad customers will become default. Correspondingly, if the credit granting policy of a financial institution is too generous, this will be exposed to high credit risk. On the other hand, type-II error (or false-alarm) defines the rate of good samples being predicted as bad; when this happens, the misclassified non-defaulters are refused and therefore, the financial institution has opportunity cost caused by the loss of those good customers. In general, type-I errors have much stronger impact on the creditor firms than type-II errors (Caouette et al 2008).

Apart from these metrics, the papers gathered in the present survey indicate that some other straightforward indices, which can be formulated from the confusion matrix, are also considered in this context (about 10% of the papers have used all or a subset of these measures). Among others, we can highlight sensitivity, specificity and precision. Sensitivity (Se) or recall is the proportion of positive cases that are correctly predicted as positive, specificity (Sp) is the proportion of negative examples that are correctly predicted as negative, and precision (Prec) or positive predictive value is defined as the proportion of cases labeled as positive. However, the use of these scores presents some apparent limitations (Hand 2012); for instance, one can achieve a sensitivity of 1 simply by predicting all the observations as positive, but at the cost of misclassifying all negative samples, thus producing a specificity of 0.

Other criteria less commonly employed in the evaluation of credit risk and bankruptcy prediction models are the mean absolute error (MAE), the root mean squared error (RMSE), the area under the ROC curve (AUC), the Gini coefficient, the Kolmogorov-Smirnov (K-S) statistic, the F-measure, and the H-measure. From these, the AUC corresponds to the most preferred score, which is usually calculated as the empirical probability that a randomly-chosen positive observation is ranked above a randomly-chosen negative example. The F-measure is a widely-used metric in information retrieval and represents the harmonic mean of sensitivity and precision, whereas the H-measure (Hand 2009) is a recently developed threshold-varying evaluation score that calculates the expected loss of the classifier (as a proportion of the maximum possible loss) under a hypothetical probability distribution of the class imbalance ratio.

Another relevant evaluation metric in the areas of finance and banking is the estimated misclassification cost (West 2000), which takes care of the unequal costs associated with making type-I or type-II errors. However, the misclassification costs are seldom available because the estimate of their values is a complex and challenging task (Lee and Chen 2005). In fact, only 5% of the papers reviewed have included the expected misclassification cost in their experimental protocol. If C_1 and C_2 denote the costs associated with type-I error

for false positives and with type-II error for false negatives respectively, then the estimated misclassification cost (Provost and Fawcett 2001) can be calculated as $Cost = C_1 \times FN + C_2 \times FP$.

Since the risk for false positives is usually much higher than that for false negatives, the assumption that the ratio of the cost C_1 to the cost C_2 is more than 2:1 is fairly realistic in this field. For example, the ratio of the misclassification cost for type-I error to the misclassification cost for type-II error in the German database was reported to be 5:1 (West 2000), which has further been taken as the ratio between the costs of both errors for other data in a number of papers (Abdou et al 2007, 2008; Abdou 2009b; Lee and Chen 2005).

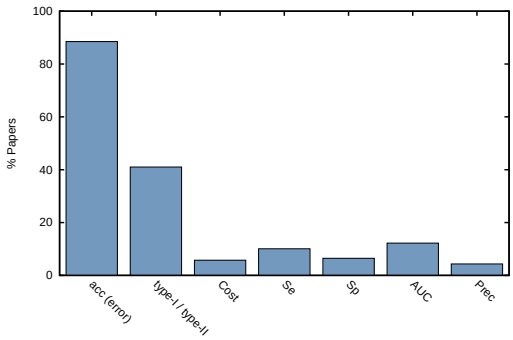

Fig. 4 Percentages of papers for each of the most commonly used metrics.

Figure 4 displays the percentages of papers that have employed each of the most typical performance metrics. As already pointed out, accuracy and error rate are the most frequently used measures in credit scoring and bankruptcy prediction. However, nearly half of these papers have also considered type-I and type-II error rates to measure the proportions of false positives and false negatives separately. AUC appears to be the third most used score in this context. Finally, although the misclassification cost is especially relevant for most applications of financial risk prediction, only a few papers have calculated this performance metric, mainly due to the difficulty of estimating the true costs associated to each type of error.

At this point, it is worth noting that a very usual problem related to credit risk and bankruptcy prediction arises when the data set is skewed, that is, the class of non-defaults (non-bankrupts) vastly outnumbers the class of defaults (bankrupts) and probably the minority class has a higher misclassification cost (Phua et al 2004; Kiefer 2009; Catal 2012). This is a very important issue that should be addressed carefully when choosing a model performance score because many metrics are biased towards the majority class and therefore they can be inappropriate for this kind of financial applications. It is surprising that the strongly biased classification accuracy is still the only measure reported in various studies, despite the voices arguing that other criteria should be used instead. In this sense, for instance, the AUC appears to be a more appropriate performance measure than accuracy for imbalanced data sets because it does not implicitly assume equal misclassification costs. However, several researchers do not take these arguments into account as can be seen in Table 5, which reports the performance metrics chosen in a number of papers with skewed databases. For instance, the negative class in the paper by Malhotra and Malhotra (2003) represents about 7% of the

whole database, but the model performance is solely evaluated by means of the prediction accuracy. Similarly, the bankrupt firms in the paper by Sun and Shenoy (2007) constitute less than 12% of the data, but the classification accuracy is the only measure included in the experiments.

Table 5 Metrics adopted for model performance evaluation in several papers with skewed databases.

Paper	%Negative	%Positive	Metrics
(Karan et al 2013)	98.3	1.7	acc, type-I, type-II, AUC, Gini, K-S, others
(Peng et al 2011)	98.1	1.9	acc, Se, Sp, Prec, AUC
(Serrano-Cinca and Gutiérrez-Nieto 2013)	97.8	2.2	acc, type-I, type-II, Se, Sp, F-measure
(Harris 2013)	97.7	2.3	acc, type-I, type-II, Se, Prec, F-measure, AUC
(Yang 2007)	96.7	3.3	acc, type-I, type-II
(Marqués et al 2012a)	95.0	5.0	acc, type-I
(García et al 2012)	95.0	5.0	acc, type-I, type-II
(Malhotra and Malhotra 2003)	93.1	6.9	acc
(Zurada and Zurada 2002)	91.1	8.9	acc, ROC curve
(Sun and Shenoy 2007)	88.6	11.4	acc
(Marinakis et al 2008)	84.5	15.5	acc, type-I, type-II, RMSE
(Zhou et al 2012)	84.2	15.8	acc, Se, Sp
(Li et al 2008)	84.0	16.0	acc, Se, Sp
(van Gestel et al 2006)	83.8	16.2	acc, Se, Sp, AUC
(Dănilă 2012)	81.7	18.3	AUC
(Chi and Tang 2006)	75.0	25.0	acc, type-I, type-II
(West et al 2005)	73.9	26.1	error, Se, Sp
(Goletsis et al 2010)	70.0	30.0	acc
(Korol 2013)	67.3	32.7	acc, type-I, type-II

6 Statistical tests of significance

It is important to take into consideration that simple superiority of a prediction model in terms of some performance score on a test set, or any other comparison based on data splitting, results naive and is not sufficient to guarantee that it certainly performs better than the rest of methods. For a complete performance evaluation, it seems pertinent to adopt some hypothesis testing in order to assert that the observed differences in performance are statistically significant, and are not merely due to random splitting effects. Statistical validation of the results has been considered for a long time an essential part of the experimental framework, but its practical use has led to much debate in several fields of science (Chow 1998; Berrar and Lozano 2013).

Choosing the right test for a specific collection of experiments depends upon several factors such as the number of data sets, the number of algorithms to be compared and the scale of measurement of the output variable (binary, nominal, interval, ordinal) (Marusteri and Bacarea 2010). On the other hand, one has also to take into account that some statistical tests are based on the assumption that the data are sampled from a normal distribution. These are the parametric tests, in contrast to the non-parametric tests which do not make assumptions about the population distribution. Although the parametric tests are, in general, more powerful than the non-parametric ones, it is not always easy to decide whether the sample comes from a normal population. In these cases, especially when the sample size is small, the use of a parametric test can be conceptually inappropriate and statistically inaccurate, and therefore it will often be preferable to apply a non-parametric procedure (Demšar 2006; García et al 2010).

Based on the review carried out, several comments can be outlined: (i) the use of statistical procedures either for determining the optimal method or for comparing the performance of different prediction models appears to be infrequent since more than 68% of papers have not reported any form of hypothesis testing; (ii) the parametric tests have been applied in nearly 18% of papers (especially the t-test with about 15%), but ignoring whether the samples hold the normality and homoscedasticity assumptions or not; (iii) approximately 13% of papers have included a non-parametric test in the experimental protocol, being the McNemar's (5.67%) and Wilcoxon's signed-ranks (3.55%) tests the two most common techniques; (iv) only three papers (Canbas et al 2005; Abdou et al 2008; Abdou 2009a) have studied the statistical difference of variances through Bartlett's, Levene's or Cochran's C tests; and (v) the post hoc tests for comparisons with a control algorithm have seldom been applied, with only seven works using the Tukey's method (Pendharkar 2005), the Nemenyi's test (García et al 2012; Marqués et al 2013; Brown and Mues 2012), the Holm's test (Hu and Chen 2011) or the Bonferroni-Dunn's procedure (Marqués et al 2012a,b).

From a practical point of view, it is possible to underline two scenarios with regard to the statistical testing of experimental results. First, the single-problem analysis involves the comparison of two or more algorithms over a unique database in terms of some metric(s). On the other hand, the multiple-problem analysis is related to the study of two or more algorithms over a number of data sets simultaneously, in terms of some performance score(s). Each of these two cases can be handled through suitable hypothesis testing methods, but we have observed that many papers simply present a matrix of tests comparing all pairs of models and then report a list of conclusions about the statistical significance for each pair. However, this kind of analysis is of little value because a proportion of the null hypotheses can be rejected by random chance (Demšar 2006).

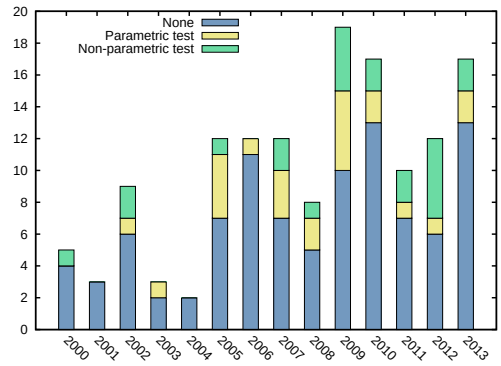

Fig. 5 Year-wise distribution of papers according to the usage of statistical tests.

The year-wise analysis of articles illustrated in Figure 5 reveals that there are not substantial differences in the application of statistical tests across the years. As can be seen, regardless of the year, a considerable majority of studies have not applied any hypothesis testing procedure. On the other hand, although the percentages of papers using parametric and non-parametric methods are very similar, the latter seems to gain some slight advantage in the last years. This suggests that the need of using appropriate tests begins now to be better understood by the research community in this field.

Despite the t-test corresponds to the most often used method for assessing the statistical significance of differences, it has been misapplied in quite a lot of studies. The most typical deficiency is that those works do not check for normality of data (e.g. Ben-David and Frank 2009; Li and Sun 2009, 2013; Martens et al 2007; Tsai and Wu 2008; Lu et al 2013; Sun and Shenoy 2007). Another problem refers to the fact that several works employ this parametric test to compare multiple algorithms (e.g. Ravisankar et al 2010; Tsai and Wu 2008; Ribeiro et al 2012), even though not being suitable to carry out this type of comparisons.

7 A straightforward experimental analysis

A couple of experimental scenarios are carried out in a more descriptive way in order to illustrate the importance of using a certain experimental methodology or another. We must remark that this study does not intend to select the best approach, but presenting an overview on how the different experimental set-ups affect the conclusions. First we analyze how the performance evaluation metrics affect the conclusions derived from the results. To this end, we use the Iranian bank database (Sabzevari et al 2007), which consists of 950 records of "good" customers and 50 samples of "bad" customers; therefore, this is an example of medium sized data set with a very strong imbalance. The second scenario is intended to show the effect of the data splitting methods over the performance of the prediction models, using in this case two benchmarking databases of different sizes: a small data set with bankruptcy information of 120 Polish companies recorded over a two-year period giving a total of 240 records (Pietruszkiewicz 2008), and the large UCSD data set with 2435 records, which corresponds to a random subset of the original database used in the 2007 Data Mining Contest organized by the University of California San Diego and Fair Isaac Corporation.

In both scenarios we have run four different prediction models: the k-nearest neighbor (k-NN) classifier with $k = 1$, the C4.5 decision tree, a support vector machine (SVM) with the linear kernel function, and a multi-layer perceptron (MLP) neural network with 10 hidden layers.

7.1 Comparing several performance evaluation metrics

The purpose of this first case study is to show the importance of using an appropriate performance evaluation measure when the data set suffers from a severe class imbalance, which has been recognized as a very common problem in the domain of credit risk and bankruptcy prediction. For this experiment, we have applied the stratified 10×5-fold cross validation resampling method and calculated some of the most widely-used metrics according to our discussion in Section 5.

Table 6 Comparison of performance evaluation measures on an imbalanced data set.

	acc	type-I	type-II	Se	Sp	AUC	Prec
1-NN	0.927	0.72	0.04	0.96	0.28	0.62	0.96
C4.5	0.939	0.84	0.02	0.98	0.16	0.60	0.96
SVM	0.950	1.00	0.00	1.00	0.00	0.50	0.95
MLP	0.940	0.85	0.02	0.98	0.15	0.72	0.96

This case represents a quite good example to illustrate that different measures can make different decisions about which algorithm is the best performing model. In addition, it allows to demonstrate that several metrics are worthless when one class is more important than the other because of the unequal costs associated with each class. Under these conditions, as already stated in Section 5, the risk for false positive (type-I) errors is usually much higher than the risk for false negative (type-II) errors.

The first observation from the results reported in Table 6 is that accuracy does not reflect the true performance of each classifier because it is biased with respect to data imbalance and proportions of correct and incorrect predictions. In fact, this measure suggests that the SVM is the best performing model, but it appears evident that its superiority comes from disregarding the minority class (type-I = 1) and assigning all samples to the majority class (type-II = 0). On the contrary, AUC, specificity and precision seem to provide a better performance evaluation in this skewed scenario since these measures propose the 1-NN classifier as the best alternative, which also corresponds to the model with the lowest type-I error and still a moderate type-II error rate. Finally, as expected from its definition, sensitivity behaves similar to accuracy and therefore it also becomes useless for performance evaluation of this strongly imbalanced data set.

7.2 Comparing various data splitting methods

In this second case study, the objective is to answer the following question: Do different data splitting methods give rise to different performance results? To this end, we compare 1-NN, C4.5, SVM and MLP on the two aforementioned benchmarking databases in terms of accuracy.

Table 7 Comparison of data splitting methods.

	Bankruptcy				UCSD			
	Holdout	5-fold CV	5×10-fold CV	LOO	Holdout	5-fold CV	5×10-fold CV	LOO
1-NN	0.764	0.750	0.763	0.767	0.753	0.761	0.758	0.764
C4.5	0.708	0.688	0.719	0.717	0.797	0.804	0.810	0.817
SVM	0.681	0.708	0.723	0.721	0.783	0.785	0.785	0.786
MLP	0.736	0.708	0.743	0.758	0.780	0.795	0.796	0.794

Table 7 summarizes the accuracy rates achieved by each of the four prediction models when using four different resampling methods with stratification: holdout (with 70/30 splits for training and test data), 5-fold cross validation, 5×10-fold cross validation and leave-one-out. Although the results seem quite similar independently of the data splitting method used, one can see that all classifiers achieve the highest accuracies with 5×10-fold cross validation and leave-one-out for the small bankruptcy database. In the case of the UCSD data set, as it has sufficient samples to form both training and test sets, all methods except the holdout approach appear to be equally valid and reliable.

8 Some final guidelines

From the discussions given in the previous sections, a number of recommendations or guidelines for researchers and practitioners who are interested in credit risk and corporate

bankruptcy prediction can be suggested. Although we do not intend to introduce stringent requirements for the design of experiments and the validation of performance results, we believe it is of utmost importance to outline a general framework with a set of key questions that leads to statistically reliable conclusions, allows for consistent comparisons among different works and supports reproducibility. Hopefully, the final guidelines given in this section will play an important role in future studies for improving rigor and objectivity of the research progress in this field.

Ideally, the empirical study of a research work should contain a mixture of benchmarking and application-oriented databases in order to profit from both views. This is especially important for this domain because the socio-economic and political dynamics of change may strongly affect the performance of the prediction models. Apart from our familiarity with the benchmarking data, which allow for an easy comparison of the performance results reported in different papers, they are also a valuable resource available for any researcher who is interested in credit scoring and corporate bankruptcy prediction. However, despite most of these data were gathered from real-life applications, it is apparent that they may represent outdated conditions of only a small portion of all possible real situations; therefore, it is not correct to generalize from the benchmarking data sets to any other data. On the contrary, application-oriented databases allow us to explore different features of the current socio-economic and political circumstances, but it may be quite difficult to access them and their knowledge is usually scarce.

The second issue to take into account refers to data splitting. Most researchers in the field apply holdout or K-fold cross validation (with $K = 5$ or $K = 10$), sometimes simply because both these are well-known and widely-used techniques. However, these methods may also present a series of limitations which should be taken into consideration carefully in order to ensure that they are certainly the most appropriate for a specific problem. In practice, it will usually be better to adopt the iterative versions of these procedures due to the generally small data set size in this kind of applications. Besides, one should take care of keeping the prior class probabilities in all partitions by applying stratification when splitting the data, thus avoiding the risk of producing a subset with no samples from some class. In the case of credit risk and bankruptcy prediction, it is very common to find small and medium sized databases with imbalanced class distribution, making even much more critical the decision of which approach to use in a proper way. Apart from these factors, it has to be noted that the choice of a data splitting method also relies on other elements such as the nature of the classifiers and the complexity of the problem.

As suggested by several researchers (Japkowicz and Shah 2011), more than a single performance score should be calculated to establish the worth of a classification model because a scalar metric cannot capture all important aspects of an algorithm. Following this recommendation in the domain of credit risk and bankruptcy prediction, the inclusion of type-I and type-II errors, probably along with other performance metrics, becomes especially important due to the unequal misclassification costs for false positives and false negatives. Another issue that should also be taken into consideration when choosing a performance evaluation measure relates to the intrinsic data characteristics because some of these may disguise the true performance of any prediction model; for instance, if the data set is skewed, which is very often in this application domain, one should discard the use of those metrics that are strongly biased towards the majority class and opt for more suitable measures.

Despite the importance of validating the performance results, we have seen that many papers lack of any statistical test of significance, while others usually apply some test without much concern about the assumptions upon which it depends. For the application of a hypothesis testing method, researchers should pay much attention to the problem they are

dealing with, that is, consider the number of prediction algorithms and the amount of experimental databases and identify the distribution of data; otherwise, the statistical test used to validate the performance results may provide misleading conclusions. The analysis carried out reveals that many papers have experimented with a unique data set, which considerably limits the number and the type of testing methods that can be applied correctly.

Finally, four simple guidelines for a more complete, robust experimental design should be kept in mind: (i) to use various databases, both benchmarking and application-oriented ones; (ii) to apply an appropriate data splitting technique according to the data set size, while preserving the prior class probabilities; (iii) to choose the scalar performance metrics depending on the data characteristics and the requirements of the problem; and (iv) to validate the results with correct statistical tests of significance taking into account the problem in hand. Although these four recommendations are indeed quite general and common to many other application domains, it is worth remarking that a large proportion of the studies analyzed have not applied them properly. On the other hand, some of these guidelines result critical in the context of credit risk and corporate bankruptcy prediction because of the special characteristics mentioned in Section 1. In addition, this is an interdisciplinary domain with researchers from very different areas, some of which are not comfortable with this kind of experimentation. These are the reasons why we believe it is still important to highlight that the experimental methodology in this field should take all these issues into consideration.

9 Conclusions

This paper has reviewed a representative sample of journal articles published between 2000 and 2013 in the context of credit risk and corporate bankruptcy prediction in order to gain insight into this subject. Unlike standard reviews in the related literature, the main objective of this work has primarily been to study and analyze the current practice in experimental design and validation of performance results, putting the emphasis on four critical components: databases, data splitting methods, performance evaluation metrics and statistical tests of significance. The relevance of this issue comes from the fact that a well-specified experimental set-up allows to reproduce the experiments by other researchers. As a by-product, however, this surveys can also be useful for new practitioners who are interested in knowing the state-of-the-art in credit and bankruptcy risk evaluation processes.

Regarding the data used in the experiments, it has been observed some shortcomings. First, the number of public databases available for experimentation is limited, thus making difficult the comparison of models between different researchers. Second, the data set size in terms of number of samples is usually small, which may increase the variance of the results (and these are more affected by chance). Third, most papers experiment with a unique database and therefore, the conclusions from these studies should be taken with caution because they may rely on the particular characteristics of such a single database.

According to the review carried out, it appears that nearly all studies have implemented some kind of data splitting, being holdout and K-fold cross validation the most frequently used procedures. However, even though the small size of many databases and the need for multiple replications, the repeated holdout and the $K_1 \times K_2$-fold cross validation have been adopted only in a very few studies. We have also found that several works do not specify which data splitting technique has been employed, thus making impossible to reproduce the experiments and acquire a complete understanding of the correctness and consistency of the

empirical results. Stratification plays an important role in resampling, but our analysis has revealed that most papers neglect this issue in the realm of credit scoring.

When analyzing the performance evaluation metrics, we have seen that a vast majority of papers have used the accuracy or the error rate, even with class imbalanced data and different misclassification costs. The problem in these cases is that the biased behavior of accuracy (and error rate) may induce misleading conclusions about the worthiness of a prediction model. Several works have tried to overcome this drawback by including also the type-I and type-II errors, which allow to assess the performance on each individual class. Another question related to model performance refers to the expected misclassification cost, which has rarely been considered in these papers in spite of the unequal costs associated to false negatives and false positives. As a final statement, it has seemed clear enough that one should choose the most appropriate performance assessment metric taking into account the particular characteristics of the data onto which a prediction model will be applied.

Finally, this survey of papers suggests that the use of statistical tests of significance is not very frequent yet. Some studies show only means and standard deviations with no hypothesis testing to conclude that one model performs better than the others, whereas some other papers apply a parametric test (mostly the t-test) without checking for normality of data. Only a few number of works have employed some non-parametric test, especially the McNemar's and Wilcoxon's methods.

Acknowledgements This work has partially been supported by the Mexican Science and Technology Council (CONACYT-Mexico) through a Postdoctoral Fellowship [223351], the Spanish Ministry of Economy under grant TIN2013-46522-P and the Generalitat Valenciana under grant PROMETEOII/2014/062.

A Credit databases

The databases used in the experiments of the studies here analyzed are presented in Table 8. For each database, we report the number of samples, the number of independent variables and the papers in which it has been employed.

Table 8: Databases used in the papers reviewed.

Database	N	D	Paper
Turkish banks	40	12	(Canbas et al 2005)
Loan payments	48	18	(Piramuthu 2006; Twala 2010)
FAME	60	12	(Wang et al 2005), (Yu et al 2008)
Tadbirpardaz	60	13	(Jouzbarkand et al 2013)
British corporations	60	5	(Becerra et al 2005)
Chinese textil sector	60	20	(Xie et al 2013)
Cinca	66	9	(van Gestel et al 2010)
Spanish non-life insurance	72	19	(Salcedo-Sanz et al 2005)
Italian bank	76	11	(Angelini et al 2008)
DATASTREAM/FT EXTEL	77	5	(Tseng and Hu 2010)
Lithuanian	100	60	(Boguslauskas and Mileris 2009; Mileris 2010)
Nanda	100	5	(van Gestel et al 2010)
Romanian small companies	105	5	(Cimpoeru 2011)
Taiwan Economic Journal	114	12	(Hu and Chen 2011)
Greek industries	118	12	(Tsakonas et al 2006)
Texas banks	118	19	(Piramuthu 2006; Twala 2010)
Tehran Stock Exchange	120	24	(Moradi et al 2013)
Moody's	129	5	(Hu and Tseng 2007; Hu 2009; van Gestel et al 2010)
Darden corporate	132	24	(Wang and Ma 2011)
Jordanian banks	140	11	(Eletter et al 2010)
Shanghai/Shenzhen Stock Exchange	153, 216	30	(Li and Sun 2009, 2011)

(Continued on next page)

Table 8 – *Continued from previous page*

Database	N	D	Paper
K&M N1	158	19	(van Gestel et al 2010)
Croatia loan association	160	31	(Bensic et al 2005)
TSEC electronic companies	160	13	(Chen 2013)
K&M N2	162	19	(van Gestel et al 2010)
Taiwan Stock Ex 1996-2004	192	7	(Cheng et al 2006)
Korea bankruptcy	195	13	(Peng et al 2011)
National Bank of Greece	210	7	(Matsatsinis 2002)
Commercial Bank of China	239	18	(Wang et al 2011; Wang and Ma 2011)
Wharton2000	240	24	(Bose 2006; Ravisankar et al 2010)
PACAP	240	24	(Chi and Tang 2006)
Pietruszkiewicz bankruptcy	240	30, 33	(García et al 2012; Marqués et al 2012a,b, 2013; Tsai and Hsu 2013)
Polish & Latin America	245	14	(Korol 2013)
Poland regional banks	295	13	(Witkowska 2006)
Ministry of Construction of China	296	16	(Liu and Zhu 2006)
Credit card	310	7	(Chen and Huang 2011)
Chinese BFP	313	5	(Li and Sun 2013)
Romanian bank	317	14	(Dănilă 2012)
S&P Compustat	200, 329	5	(Pendharkar 2005; West et al 2005)
B&K A	342	5	(van Gestel et al 2010)
National Bank of Belgium	366	11	(Cielen et al 2004)
Israel institution	390	–	(Ben-David and Frank 2009)
Bene-C	422	15	(Martens et al 2007)
Norwegian Register of Bankruptcies	422	28	(Lensberg et al 2006)
Benelux	422	40	(van Gestel et al 2010, 2006)
FDIC 1991-1992	480	93	(Zhao et al 2009)
Atiya	481	63	(van Gestel et al 2010)
Taipei housing loan	510	18	(Lee and Chen 2005)
Manufacturing firms	528	9	(Shin and Lee 2002)
Egyptian bank	581	12, 20	(Abdou et al 2007, 2008; Abdou 2009a)
Slovenian bank	581	21	(Sušteršič et al 2009)
Russian banks	588	12	(Lanine and Vennet 2006)
Japanese	653	15	(García et al 2012; Hamadani 2013; Marqués et al 2012a,b, 2013; Nanni and Lumini 2009; Peng et al 2008, 2011; Tsai and Wu 2008; Tsai 2009; Tsai and Cheng 2012; Tsai and Hsu 2013; Wang et al 2005; Yu et al 2008, 2009)
Australian	690	14	(Baesens et al 2003; Brown and Mues 2012; Chen and Huang 2003; Chen and Li 2010; Elsayad 2010; Flórez-López 2010; García et al 2012; Goletsis et al 2010; Hamadani 2013; Hoffmann et al 2007; Hsieh 2005; Huang et al 2006, 2007; Kotsiantis 2007; Lacerda et al 2005; Li and Sun 2011; Marqués et al 2012a,b, 2013; Martens et al 2007; Nanni and Lumini 2009; Ong et al 2005; Peng et al 2008, 2011; Ping and Yongheng 2011; Piramuthu 2006; Siami et al 2013; Tsai and Wu 2008; Tsai 2009; Tsai and Cheng 2012; Tsai and Hsu 2013; Twala 2010; Wang and Huang 2009; Wang et al 2011, 2012; West 2000; West et al 2005; Zhang et al 2010; Zhou et al 2009, 2011)
Croatian bank	825	21	(Pervan and Kuvek 2013)
European companies	1000	23	(Callejón et al 2013)
German	1000	20, 24	(Baesens et al 2003; Brown and Mues 2012; Chen and Li 2010; García et al 2012; Goletsis et al 2010; Hamadani 2013; Hoffmann et al 2007; Hsieh 2005; Huang et al 2006, 2007; Khashman 2010; Kim and Sohn 2004; Koh et al 2006; Kotsiantis 2007; Laha 2007; Li and Sun 2011; Lu et al 2013; Marqués et al 2012a,b, 2013; Nanni and Lumini 2009; Ong et al 2005; Peng et al 2008, 2011; Ping and Yongheng 2011; Piramuthu 2006; Siami et al 2013; Tsai and Wu 2008; Tsai 2009; Tsai and Cheng 2012; Tsai and Hsu 2013; Twala 2010; Wang and Huang 2009; Wang et al 2011, 2012; West 2000; West et al 2005; Yu et al 2009; Zhang et al 2010; Zhou et al 2009, 2011)
Iranian bank	1000	27	(García et al 2012; Marqués et al 2012a,b, 2013)

(Continued on next page)

Table 8 – *Continued from previous page*

Database	*N*	*D*	Paper
Korea manufacturers	1000	56	(Cho et al 2010)
Yobas	1001	14	(Yobas et al 2000)
Korean enterprises	1009	–	(Hong 2009)
Credit unions	1078	6	(Malhotra and Malhotra 2002, 2003)
Belgium bank	1102	21	(Daubie et al 2002)
US firms	1160	6	(Atiya 2001)
Spanish SABI	1180	16	(Alfaro et al 2008)
Behavioural	1197	60	(Brown and Mues 2012)
DIANE	1200	30	(Chen et al 2011; Ribeiro et al 2012)
Thomas	1225	12, 14	(García et al 2012; Goletsis et al 2010; Wang et al 2005; Yu et al 2009)
Egypt banks	1262	15	(Abdou 2009b)
Commercial Bank of Greece	1411	16	(Marinakis et al 2008)
Finnish enterprises	1500	23	(Kaski et al 2001)
Korea credit guarantee	1888	11	(Min and Lee 2005)
UK4	1980	19	(Baesens et al 2003)
China local bank	2000	15	(Chen et al 2009)
Industrial Bank of Korea	2144	13	(Park and Han 2002)
NSW credit unions	2144	30	(Tan and Dihardjo 2001)
Korean firms	2400	8	(Ahn et al 2000)
UCSD	2435	38, 40	(García et al 2012; Marqués et al 2012a,b, 2013; Tsai 2009; Tsai and Cheng 2012)
Korea bank	2670	15	(Ahn and Kim 2009)
Bene1-Brown	2974	27	(Brown and Mues 2012)
Taiwan bank	3000	10	(Tsai and Chen 2010)
Bene1	3123	28, 33	(Baesens et al 2003; Hoffmann et al 2002, 2007)
Money lending	3364	12	(Zurada and Zurada 2002)
German corporates	3599	98	(Fritz and Hosemann 2000)
UK3	3960	19	(Baesens et al 2003)
Mexican bank	4000	24	(Galindo and Tamayo 2000)
Greek bank	5340	11	(Antonakis and Sfakianakis 2009)
China commercial bank	5456	13	(Peng et al 2011)
Taipei local bank	6000	9	(Lee et al 2002, 2006)
German creditor 1	6000	102	(Yang 2007)
US bank	6000	66	(Li et al 2008; Peng et al 2008, 2011; Zhou et al 2011)
Turkish retail stores	6304	20	(Karan et al 2013)
Italian CERVED	7113	10	(Ciampi and Gordini 2013)
Bene2	7190	27, 33	(Baesens et al 2003; Brown and Mues 2012; Hoffmann et al 2007)
North America Wharton	7728	27	(Zhou et al 2012)
NASDAQ	7822	20	(Sun and Shenoy 2007)
FDIC 2008-2011	8013	17	(Serrano-Cinca and Gutiérrez-Nieto 2013)
UK1	9360	16	(Baesens et al 2003)
IBM Global Finance/Experian Italy	9730	30	(Paleologo et al 2010)
UK2	11700	16	(Baesens et al 2003)
US insurance corporation	18875	103	(Peng et al 2011)
Barbados credit union	21620	19	(Harris 2013)
Taiwan credit card	25000	23	(Yeh and Lien 2009)
Credit card	25000	34	(Bellotti and Crook 2009)
German creditor 2	27633	68	(Yang 2007)
Wharton JPNBD	36636	10	(Zhou 2013)
German credit insurance	38283	72	(Liu and Schumann 2005)
Wharton USABD	86129	10	(Zhou 2013)
Experian UK	88789	39	(Finlay 2011)
UPL	92258	20	(Pavlidis et al 2012)
UCSD-2	106777	40	(Tsai and Hsu 2013)
US credit card	212742	75	(Im et al 2012)
US consumer data	–	–	(Khandani et al 2010)
Ukraine bank	–	–	(Pavlenko and Chernyak 2010)

References

Abdou H, Pointon J (2011) Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management 18(2–3):59–88

Abdou H, Pointon J, El-Masry A (2008) Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert Systems with Applications 35(3):1275–1292

Abdou HA (2009a) An evaluation of alternative scoring models in private banking. Journal of Risk Finance 10(1):38–53

Abdou HA (2009b) Genetic programming for credit scoring: The case of Egyptian public sector banks. Expert Systems with Applications 36(9):11,402–11,417

Abdou HA, El-Masry A, Pointon J, Abdou H, El-Masry A, Pointon J (2007) On the applicability of credit scoring models in Egyptian banks. Banks and Bank Systems 2(1):4–20

Ahn BS, Cho SS, Kim CY (2000) The integrated methodology of rough set theory and artificial neural network for business failure prediction. Expert Systems with Applications 18(2):65–74

Ahn H, Kim KJ (2009) Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. Applied Soft Computing 9(2):599–607

Alfaro E, García N, Gámez M, Elizondo D (2008) Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. Decision Support Systems 45(1):110–122

Alpaydin E (2010) Introduction to Machine Learning. MIT Press, Cambridge, MA

Angelini E, di Tollo G, Roli A (2008) A neural network approach for credit risk evaluation. The Quarterly Review of Economics and Finance 48(4):733–755

Antonakis AC, Sfakianakis ME (2009) Assessing naïve Bayes as a method for screening credit applicants. Journal of Applied Statistics 36(5):537–545

Atiya AF (2001) Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Trans on Neural Networks 12(4):929–935

Bache K, Lichman M (2013) UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, URL `http://archive.ics.uci.edu/ml`

Baesens B, van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society 54(6):627–635

Becerra VM, Galvao RKH, Abou-Seads M (2005) Neural and wavelet network models for financial distress classification. Data Mining and Knowledge Discovery 11(1):35–55

Bellotti T, Crook J (2009) Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications 36(2):3302–3308

Ben-David A, Frank E (2009) Accuracy of machine learning models versus "hand crafted" expert systems – a credit scoring case study. Expert Systems with Applications 36(3):5264–5271

Bensic M, Sarlija N, Zekic-Susac M (2005) Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intelligent Systems in Accounting, Finance and Management 13(3):133–150

Berrar D, Lozano JA (2013) Significance tests or confidence intervals: which are preferable for the comparison of classifiers? Journal of Experimental & Theoretical Artificial Intelligence 25(2):189–206

Bischl B, Mersmann O, Trautmann H, Weihs C (2012) Resampling methods for meta-model validation with recommendations for evolutionary computation. Evol Comput 20(2):249–275

Boguslauskas V, Mileris R (2009) Estimation of credit risk by artificial neural networks models. Economics of Engineering Decisions 4:7–14

Bose I (2006) Deciding the financial health of dot-coms using rough sets. Information & Management 43(7):836–846

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

Callejón AM, Casado AM, Fernández MA, Peláez JI (2013) A system of insolvency prediction for industrial companies using a financial alternative model with neural networks. International Journal of Computational Intelligence Systems 6(1):29–37

Canbas S, Cabuk A, Kilic SB (2005) Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case. European Journal of Operational Research 166(2):528–546

Caouette JB, Altman EI, Narayanan P, Nimmo R (2008) Managing Credit Risk: The Great Challenge for Global Financial Markets. Wiley, Hoboken, NJ

Catal C (2012) Performance evaluation metrics for software fault prediction studies. Acta Polytechnica Hungarica 9(4):193–206

Chen FL, Li FC (2010) Combination of feature selection approaches with SVM in credit scoring. Expert Systems with Applications 37(7):4902–4909

Chen MC, Huang SH (2003) Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications 24(4):433–441
Chen MY (2013) A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering. Information Sciences 220:180–195
Chen N, Ribeiro B, Vieira AS, Duarte J, Neves JC (2011) A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. Expert Systems with Applications 38(10):12,939–12,945
Chen SC, Huang MY (2011) Constructing credit auditing and control management model with data mining technique. Expert Systems with Applications 38(5):5359–5365
Chen W, Ma C, Ma L (2009) Mining the customer credit using hybrid support vector machine technique. Expert Systems with Applications 36(4):7611–7616
Cheng CB, Chen CL, Fu CJ (2006) Financial distress prediction by a radial basis function network with logit analysis learning. Computers & Mathematics with Applications 51(3–4):579–588
Chi LC, Tang TC (2006) Bankruptcy prediction: Application of logit analysis in export credit risks. Australian Journal of Management 31(1):17–28
Cho S, Hong H, Ha BC (2010) A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the mahalanobis distance: For bankruptcy prediction. Expert Systems with Applications 37(4):3482–3488
Chow SL (1998) Precis of statistical significance: rationale, validity, and utility. Behavioral and Brain Sciences 21(2):169–239
Ciampi F, Gordini N (2013) Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of Italian small enterprises. Journal of Small Business Management 51(1):23–45
Cielen A, Peeters L, Vanhoof K (2004) Bankruptcy prediction using a data envelopment analysis. European Journal of Operational Research 154(2):526–532
Cimpoeru SS (2011) Neural networks and their application in credit risk assessment. evidence from the Romanian market. Technological and Economic Development of Economy 17(3):519–534
Cohen PR (1995) Empirical Methods for Artificial Intelligence. MIT Press, Cambridge, MA
Crook JN, Edelman DB, Thomas LC (2007) Recent developments in consumer credit risk assessment. European Journal of Operational Research 183(3):1447–1465
Dănilă OM (2012) Credit risk assessment under Basel Accords. Theoretical and Applied Economics 18(3):77–90
Daubie M, Levecq P, Meskens N (2002) A comparison of the rough sets and recursive partitioning induction approaches: An application to commercial loans. International Transactions in Operational Research 9(5):681–694
Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7(1):1–30
Eletter SF, Yaseen SG, Elrefae GA (2010) Neuro-based artificial intelligence model for loan decisions. American Journal of Economics and Business Administration 2(1):27–34
Elsayad AM (2010) Implementing automated prediction systems for credit scoring. ICGST International Journal on Automatic Control and Systems Engineering 10(1):11–19
Finlay S (2011) Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research 210(2):368–378
Flórez-López R (2010) Effects of missing data in credit risk scoring. a comparative analysis of methods to achieve robustness in the absence of sufficient data. Journal of the Operational Research Society 61(3):486–501
Forman G, Scholz M (2010) Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. SIGKDD Explorations Newsletters 12(1):49–57
Fritz S, Hosemann D (2000) Restructuring the credit process: Behaviour scoring for German corporates. International Journal of Intelligent Systems in Accounting, Finance & Management 9(1):9–21
Galindo J, Tamayo P (2000) Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Computational Economics 15(1–2):107–143
García S, Fernández A, Luengo J, Herrera F (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciencies 180(10):2044–2064
García V, Marqués AI, Sánchez JS (2012) On the use of data filtering techniques for credit risk prediction with instance-based models. Expert Systems with Applications 39(18):13,267–13,276
van Gestel T, Baesens B, Suykens JAK, den Poel DV, Baestaens DE, Willekens M (2006) Bayesian kernel based classification for financial distress detection. European Journal of Operational Research 172(3):979–1003
van Gestel T, Baesens B, Martens D (2010) From linear to non-linear kernel based classifiers for bankruptcy prediction. Neurocomputing 73(16–18):2955–2970

Goletsis Y, Exarchos TP, Katsis CD (2010) Credit scoring using an Ant mining approach. Human Systems Management 29(2):79–88
Hamadani AZ (2013) An integrated genetic-based model of naive Bayes networks for credit scoring. International Journal of Artificial Intelligence & Applications 4(1):85–103
Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the roc curve. Machine Learning 77(1):103–123
Hand DJ (2012) Assessing the performance of classification methods. International Statistical Review 80(3):400–414
Harris T (2013) Quantitative credit risk assessment using support vector machines: Broad versus narrow default definitions. Expert Systems with Applications DOI 10.1016/j.eswa.2013.01.044
Hoffmann F, Baesens B, Martens J, Put F, Vanthienen J (2002) Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring. International Journal of Intelligent Systems 17(11):1067–1083
Hoffmann F, Baesens B, Mues C, van Gestel T, Vanthienen J (2007) Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. European Journal of Operational Research 177(1):540–555
Hong CS (2009) Optimal threshold from ROC and CAP curves. Communications in Statistics – Simulation and Computation 38(10):2060–2072
Horcher KA (2005) Essentials of Financial Risk Management. Wiley, Hoboken, NJ
Hsieh NC (2005) Hybrid mining approach in the design of credit scoring models. Expert Systems with Applications 28(4):655–665
Hu YC (2009) Bankruptcy prediction using ELECTRE-based single-layer perceptron. Neurocomputing 72(13–15):3150–3157
Hu YC, Chen CJ (2011) A PROMETHEE-based classification method using concordance and discordance relations and its application to bankruptcy prediction. Information Sciences 181(22):4959–4968
Hu YC, Tseng FM (2007) Functional-link net with fuzzy integral for bankruptcy prediction. Neurocomputing 70(16–18):2959–2968
Huang CL, Chen MC, Wang CJ (2007) Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications 33(4):847–856
Huang JJ, Tzeng GH, Ong CS (2006) Two-stage genetic programming (2SGP) for the credit scoring model. Applied Mathematics and Computation 174(2):1039–1053
Hughes G (1968) On the mean accuracy of statistical pattern recognizers. IEEE Trans on Information Theory 14(1):55–63
Im JK, Apley DW, Qi C, Shan X (2012) A time-dependent proportional hazards survival model for credit risk analysis. Journal of the Operational Research Society 63(3):306–321
Japkowicz N, Shah M (2011) Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, New York, NY
Jouzbarkand M, Keivani FS, Khodadadi M, Fahim SRSN, Aghajani V (2013) Bankruptcy prediction model by Ohlson and Shirata models in Tehran stock exchange. World Applied Sciences Journal 21(2):152–156
Karan MB, Ulucan A, Kaya M (2013) Credit risk estimation using payment history data: A comparative study of Turkish retail stores. Central European Journal of Operations Research 21:1–16
Kaski S, Sinkkonen J, Peltonen J (2001) Bankruptcy analysis with selforganizing maps in learning metrics. IEEE Trans on Neural Networks 12(4):936–947
Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance 34(11):2767–2787
Khashman A (2010) Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications 37(9):6233–6239
Kiefer NM (2009) Default estimation for low-default portfolios. Journal of Empirical Finance 16(1):164–173
Kim YS, Sohn SY (2004) Managing loan customers using misclassification patterns of credit scoring model. Expert Systems with Applications 26(4):567–573
Koh HC, Tan WC, Goh CP (2006) A two-step method to construct credit scoring models with data mining techniques. International Journal of Business and Information 1(1):96–118
Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. 14th International Joint Conference on Artificial intelligence, vol 2, Montreal, Canada, pp 1137–1143
Korol T (2013) Early warning models against bankruptcy risk for Central European and Latin American enterprises. Economic Modelling 31(1):22–30
Kotsiantis S (2007) Credit risk analysis using a hybrid data mining model. International Journal of Intelligent Systems Technologies and Applications 2(4):345–356
Lacerda E, Carvalho ACPLF, Braga AP, Ludermir TB (2005) Evolutionary radial basis functions for credit assessment. Applied Intelligence 22(3):167–181

Laha A (2007) Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring. Advanced Engineering Informatics 21(3):281–291
Lanine G, Vennet RV (2006) Failure prediction in the Russian bank sector with logit and trait recognition models. Expert Systems with Applications 30(3):463–478
Lee TS, Chen IF (2005) A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications 28(4):743–752
Lee TS, Chiu CC, Lu CJ, Chen IF (2002) Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications 23(3):245–254
Lee TS, Chiu CC, Chou YC, Lu CJ (2006) Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis 50(4):1113–1130
Lensberg T, Eilifsen A, McKee TE (2006) Bankruptcy theory development and classification via genetic programming. European Journal of Operational Research 169(2):677–697
Li A, Shi Y, He J (2008) MCLP-based methods for improving "bad" catching rate in credit cardholder behavior analysis-based methods for improving "bad" catching rate in credit cardholder behavior analysis. Applied Soft Computing 8(3):1259–1265
Li H, Sun J (2009) Gaussian case-based reasoning for business failure prediction with empirical data in China. Information Sciences 179(1–2):89–108
Li H, Sun J (2011) Principal component case-based reasoning ensemble for business failure prediction. Information & Management 48(6):220–227
Li H, Sun J (2013) Predicting business failure using an RSF-based case-based reasoning ensemble forecasting method. Journal of Forecasting 32(2):180–192
Lin WY, Hu YH, Tsai CF (2012) Machine learning in financial crisis: A survey. IEEE Trans on Systems, Man, and Cybernetics–Part C: Applications and Reviews 42(4):421–436
Liu G, Zhu Y (2006) Credit assessment of contractors: A rough set method. Tsinghua Science & Technology 11(3):357–362
Liu Y, Schumann M (2005) Data mining feature selection for credit scoring models. Journal of the Operational Research Society 56(9):1099–1108
Lu H, Liyan H, Hongwei Z (2013) Credit scoring model hybridizing artificial intelligence with logistic regression. Journal of Networks 8(1):253–261
Malhotra R, Malhotra DK (2002) Differentiating between good credits and bad credits using neuro-fuzzy systems. European Journal of Operational Research 136(1):190–211
Malhotra R, Malhotra DK (2003) Evaluating consumer loans using neural networks. Omega 31(2):83–96
Marinakis Y, Marinaki M, Doumpos M, Matsatsinis N, Zopounidis C (2008) Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment. Journal of Global Optimization 42(2):279–293
Marqués AI, García V, Sánchez JS (2012a) Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications 39(11):10,244–10,250
Marqués AI, García V, Sánchez JS (2012b) Two-level classifier ensembles for credit risk assessment. Expert Systems with Applications 39(12):10,916–10,922
Marqués AI, García V, Sánchez JS (2013) On the suitability of resampling techniques for the class imbalance problem in credit scoring. Journal of the Operational Research Society 64(7):1060–1070
Martens D, Baesens B, van Gestel T, Vanthienen J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research 183(3):1466–1476
Marusteri M, Bacarea V (2010) Comparing groups for statistical differences: how to choose the right statistical test? Biochemia Medica 20(1):15–32
Matsatsinis NF (2002) CCAS: An intelligent decision support system for credit card assessment. Journal of Multi-Criteria Decision Analysis 11(4–5):213–235
Mileris R (2010) Estimation of loan applicants default probability applying discriminant analysis and simple Bayesian classifier. Economics and Management 15:1078–1084
Min JH, Lee YC (2005) Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 28(4):603–614
Moradi M, Salehi M, Ghorgani ME, Yazdi HS (2013) Financial distress prediction of Iranian companies using data mining techniques. Organizacija 46(1):20–27
Nagy G (2004) Classifiers that improve with use. In: Proc. Conference on Pattern Recognition and Multimedia, Tokyo, Japan, pp 79–86
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
Ong CS, Huang JJ, Tzeng GH (2005) Building credit scoring models using genetic programming. Expert Systems with Applications 29(1):41–47

Paleologo G, Elisseeff A, Antonini G (2010) Subagging for credit scoring models. European Journal of Operational Research 201(2):490–499
Park CS, Han I (2002) A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. Expert Systems with Applications 23(3):255–264
Pavlenko T, Chernyak O (2010) Credit risk modeling using bayesian networks. International Journal of Intelligent Systems 25:326–344
Pavlidis NG, Tasoulis DK, Adams NM, Hand DJ (2012) Adaptive consumer credit classification. Journal of the Operational Research Society 63(12):1645–1654
Pendharkar PC (2005) A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. Computers & Operations Research 32(10):2561–2582
Peng Y, Kou G, Shi Y, Chen Z (2008) A multi-criteria convex quadratic programming model for credit data analysis. Decision Support Systems 44(4):1016–1030
Peng Y, Wang G, Kou G, Shi Y (2011) An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing 11(2):2906–2915
Pervan I, Kuvek T (2013) The relative importance of financial ratios and nonfinancial variables in predicting of insolvency. Croatian Operational Research Review 4(1):187–197
Phua C, Alahakoon D, Lee V (2004) Minority report in fraud detection: Classification of skewed data. SIGKDD Explorations Newsletter 6(1):50–59
Pietruszkiewicz W (2008) Dynamical systems and nonlinear Kalman filtering applied in classification. In: Proc. of 7th IEEE International Conference on Cybernetic Intelligent Systems, London, UK, pp 263–268
Ping Y, Yongheng L (2011) Neighborhood rough set and SVM based hybrid credit scoring classifier. Expert Systems with Applications 38(9):11,300–11,304
Piramuthu S (2006) On preprocessing data for financial credit risk evaluation. Expert Systems with Applications 30(3):489–497
Provost FJ, Fawcett T (2001) Robust classification for imprecise environments. Machine Learning 42(3):203–231
Raeder T, Forman G, Chawla N (2012) Learning from imbalanced data: Evaluation matters. In: Holmes DE, Jain LC (eds) Data Mining: Foundations and Intelligent Paradigms, Springer-Verlag, Berlin Heidelberg, pp 315–331
Ravi Kumar P, Ravi V (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. European Journal of Operational Research 180(1):1–28
Ravisankar P, Ravi V, Bose I (2010) Failure prediction of dotcom companies using neural network-genetic programming hybrids. Information Sciences 180(8):1257–1267
Ribeiro B, Silva C, Chen N, Vieira AS, Neves JC (2012) Enhanced default risk models with SVM+. Expert Systems with Applications 39(11):10,140–10,152
Sabzevari H, Soleymani M, Noorbakhsh E (2007) A comparison between statistical and data mining methods for credit scoring in case of limited available data. In: Proc. of the 3rd CRC Credit Scoring Conference, Edinburgh, UK
Sadatrasoul SM, Gholamian MR, Siami M, Hajimohammadi Z (2013) Credit scoring in banks and financial institutions via data mining techniques: A literature review. Journal of AI and Data Mining 1(2):119–129
Salcedo-Sanz S, Fernández-Villacañas JL, Segovia-Vargas MJ, Bousoño-Calzón C (2005) Genetic programming for the prediction of insolvency in non-life insurance companies. Computers & Operations Research 32(4):749–765
Sechidis K, Tsoumakas G, Vlahavas I (2011) On the stratification of multi-label data. In: Proc. European Conference on Machine Learning and Knowledge Discovery in Databases, vol 2, Athens, Greece, pp 145–158
Serrano-Cinca C, Gutiérrez-Nieto B (2013) Partial least square discriminant analysis for bankruptcy prediction. Decision Support Systems 54(3):1245–1255
Shin KS, Lee YJ (2002) A genetic algorithm application in bankruptcy prediction modeling. Expert Systems with Applications 23(3):321–328
Siami M, Gholamian MR, Basiri J (2013) An application of locally linear model tree algorithm with combination of feature selection in credit scoring. International Journal of Systems Science DOI 10.1080/00207721.2013.767395
Staples M, Niazi M (2007) Experiences using systematic review guidelines. Journal of Systems and Software 80(9):1425–1437
Sun L, Shenoy PP (2007) Using Bayesian networks for bankruptcy prediction: Some methodological issues. European Journal of Operational Research 180(2):738–753
Sušteršič M, Mramor D, Zupan J (2009) Consumer credit scoring models with limited data. Expert Systems with Applications 36(3):4736–4744

Tan CNW, Dihardjo H (2001) A study of using artificial neural networks to develop an early warning predictor for credit union financial distress with comparison to the probit model. Managerial Finance 27(4):56–77
Thomas LC, Edelman DB, Crook JN (2002) Credit Scoring and Its Applications. SIAM, Philadelphia, PA
Tsai CF (2009) Feature selection in bankruptcy prediction. Knowledge-Based Systems 22(2):120–127
Tsai CF, Chen ML (2010) Credit rating by hybrid machine learning techniques. Applied Soft Computing 10(2):374–380
Tsai CF, Cheng KC (2012) Simple instance selection for bankruptcy prediction. Knowledge-Based Systems 27:333–342
Tsai CF, Hsu YF (2013) A meta-learning framework for bankruptcy prediction. Journal of Forecasting 32(2):167–179
Tsai CF, Wu JW (2008) Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications 34(4):2639–2649
Tsakonas A, Dounias G, Doumpos M, Zopounidis C (2006) Bankruptcy prediction with neural logic networks by means of grammar-guided genetic programming. Expert Systems with Applications 30(3):449–461
Tseng FM, Hu YC (2010) Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. Expert Systems with Applications 37(3):1846–1853
Twala B (2010) Multiple classifier application to credit risk assessment. Expert Systems with Applications 37(4):3326–3336
Verikas A, Kalsyte Z, Bacauskiene M, Gelzinis A (2010) Hybrid and ensemble-based soft computing techniques in bankruptcy prediction : A survey. Soft Computing - A Fusion of Foundations, Methodologies and Applications 14(9):995–1010
Wang CM, Huang YF (2009) Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. Expert Systems with Applications 36(3):5900–5908
Wang G, Ma J (2011) Study of corporate credit risk prediction based on integrating boosting and random subspace. Expert Systems with Applications 38(11):13,871–13,878
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
Wang G, Ma J, Huang L, Xu K (2012) Two credit scoring models based on dual strategy ensemble trees. Knowledge-Based Systems 26:61–68
Wang Y, Wang S, Lai KK (2005) A new fuzzy support vector machine to evaluate credit risk. IEEE Trans on Fuzzy Systems 13(6):820–831
West D (2000) Neural network credit scoring models. Computers & Operations Research 27(11–12):1131–1152
West D, Dellana S, Qian J (2005) Neural network ensemble strategies for financial decision applications. Computers & Operations Research 32(10):2543–2559
Witkowska D (2006) Discrete choice model application to the credit risk evaluation. International Advances in Economic Research 12(1):33–42
Xie G, Zhao Y, Jiang M, Zhang N (2013) A novel ensemble learning approach for corporate financial distress forecasting in fashion and textiles supply chains. Mathematical Problems in Engineering pp 1–9
Yang Y (2007) Adaptive credit scoring with kernel learning methods. European Journal of Operational Research 183(3):1521–1536
Yeh IC, Lien CH (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36(2):2473–2480
Yobas MB, Crook JN, Ross P (2000) Credit scoring using neural and evolutionary techniques. IMA Journal of Management Mathematics 11(2):111–125
Yu L, Wang S, Lai KK (2008) Credit risk assessment with a multistage neural network ensemble learning approach. Expert Systems with Applications 34(2):1434–1444
Yu L, Wang S, Lai KK (2009) An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. European Journal of Operational Research 195(3):942–959
Zhang D, Zhou X, Leung SCH, Zheng J (2010) Vertical bagging decision trees model for credit scoring. Expert Systems with Applications 37(12):7838–7843
Zhao H, Sinha A, Ge W (2009) Effects of feature construction on classification performance: An empirical study in bank failure prediction. Expert Systems with Applications 36(2):2633–2644
Zhou L (2013) Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. Knowledge-Based Systems 41:16–25
Zhou L, Lai KK, Yu L (2009) Credit scoring using support vector machines with direct search for parameters selection. Soft Computing 13(2):149–155
Zhou L, Lai KK, Yen J (2012) Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation. International Journal of Systems Science pp 1–13

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

Zurada J, Zurada M (2002) How secure are "good loans": Validating loan-granting decisions and predicting default rates on consumer loans. Review of Business Information Systems 6(3):65–84