ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 June: 21(2): 173~190 http://dx.doi.org/10.13088/jiis.2015.21.2.173

A Hybrid Under-sampling Approach for Better Bankruptcy Prediction*

Taehoon Kim
Master's Candidate, Graduate School of Business IT, Kookmin University
(kth8408@naver.com)

Hyunchul Ahn
Associate Professor, Graduate School of Business IT, Kookmin University
(hcahn@kookmin.ac.kr)

The purpose of this study is to improve bankruptcy prediction models by using a novel hybrid under-sampling approach. Most prior studies have tried to enhance the accuracy of bankruptcy prediction models by improving the classification methods involved. In contrast, we focus on appropriate data preprocessing as a means of enhancing accuracy. In particular, we aim to develop an effective sampling approach for bankruptcy prediction, since most prediction models suffer from class imbalance problems. The approach proposed in this study is a hybrid under-sampling method that combines the k-Reverse Nearest Neighbor (k-RNN) and one-class support vector machine (OCSVM) approaches. k-RNN can effectively eliminate outliers, while OCSVM contributes to the selection of informative training samples from majority class data. To validate our proposed approach, we have applied it to data from H Bank's non-external auditing companies in Korea, and compared the performances of the classifiers with the proposed under-sampling and random sampling data. The empirical results show that the proposed under-sampling approach generally improves the accuracy of classifiers, such as logistic regression, discriminant analysis, decision tree, and support vector machines. They also show that the proposed under-sampling approach reduces the risk of false negative errors, which lead to higher misclassification costs.

주제어 : Bankruptcy Prediction, Under-sampling, k-Reverse Nearest Neighbor, One-class Support Vector Machine, Classification

Received : May 20, 2015 Revised : June 15, 2015 Accepted : June 16, 2015
Type of submission : Fast Track Corresponding Author : Hyunchul Ahn

1. Introduction

When a company goes bankrupt, it can lead to huge social and economic losses. In general, the economic value of a company is measured by asset value, service value, human value, and symbolic value. However, when a bankruptcy occurs, the rated value of the company is based only on material asset value, and so the economic value of the company is reduced. This reduction in economic value leads to a corresponding reduction of creditor and shareholder wealth. In addition, the

* This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A5A2A03064791)

bankruptcy of businesses is linked to the performance degradation of associated companies and partners through effects such as the degradation of industrial productivity, and can result in unemployment. If a bankrupt company is then acquired by a foreign company, core technology is spilled, which may reduce national industrial competitiveness. Therefore, to avoid such problems and minimize damage, it is necessary to accurately predict bankruptcy (Kim et al., 2011). Bankruptcy prediction helps financial institutions or companies to minimize the predictable losses of stakeholders, such as shareholders, creditors, suppliers, and workers, by providing appropriate information (Bellovary et al., 2007).

Bankruptcy prediction involves typical binary classification approaches that classify bankrupt and non-bankrupt companies. Early studies, from the 1960s to 1980s, tested statistical techniques as classifiers (Altman, 1968; Deakin, 1974; Ohlson, 1980). Since the 1990s, substantial research into the use of artificial intelligence (AI) and data mining techniques for bankruptcy prediction has been carried out. The studies involved have tested decision trees, artificial neural networks, support vector machines, and other AI approaches, or their hybrid algorithms, as classification techniques (Kumar and Ravi, 2007). In other words, the vast majority of prior studies on bankruptcy prediction have focused on searching for better classification techniques. However, it is also important to employ a better data preprocessing procedure to improve the accuracy of bankruptcy prediction models.

With this background in mind, we pay attention to appropriate data preprocessing as a means of enhancing accuracy. In particular, we aim to develop effective sampling approaches for bankruptcy prediction, since most bankruptcy prediction models suffer from class imbalance problems. The approach proposed in this study is a hybrid under-sampling method that combines the k-RNN and OCSVM methods. k-RNN can effectively eliminate outliers, while OCSVM contributes to the selection of informative training samples from data of the majority class.

The rest of this paper is organized as follows. In section 2, we briefly review prior studies related to bankruptcy prediction and imbalanced data handling. We then introduce our research model－a novel hybrid under-sampling approach－in section 3. An experiment in empirical validation and its results are presented in section 4. Finally, we suggest conclusions and limitations of the study in the final section.

2. Literature Review

2.1. Bankruptcy Prediction

Bankruptcy prediction is important for both creditors and investors. Financial institutions that loan money to borrowers want to minimize the risk of their borrowers defaulting (declaring bankruptcy). When investors invest in a company, they also hope to avoid the bankruptcy of the

company, as it could lead to the loss of their entire investment. Thus, it is important for creditors and investors to accurately predict corporate bankruptcy (Zhou, 2013). To address these needs within the financial industry, a huge amount of studies have been conducted on bankruptcy prediction.

Existing literature on bankruptcy prediction dates back to the 1930s. The early studies, dating up to the mid-1960s, analyzed univariate ratio analysis to predict future bankruptcy (Bellovary et al., 2007). The multivariate approach, which has remained mainstream until today, was first applied by Altman (1968). Later studies, from the late 1960s until the 1980s, used statistical methods such as discriminant analysis (DA) and logit and probit models to predict corporate bankruptcy (Altman, 1968; Deakin, 1974; Ohlson, 1980).

However, as time has passed, advancements and technological developments have made other methods more prominent (i.e., artificial intelligence and data mining techniques). For example, Odom and Sharda (1990) have analyzed financial ratios by using DA and the artificial neural network (ANN), and have compared their prediction performances. Their results show that the prediction performance of ANN is superior to DA.

Tam and Kiang (1992) have compared and analyzed two artificial intelligence techniques－ANN and decision trees－as well as three statistical techniques－DA, logit analysis, and k-nearest neighbor method (k-NN)－for predicting the bankruptcy of banks. Their empirical results showed that ANN was superior to other techniques.

A recent study by Tai and Shin(2009) also proposed ANN as the classifier for bankruptcy prediction. However, they proposed to use genetic algorithm(GA) based normalization approach in order to improve prediction accuracy.

Modified versions of ANN have also been used to predict corporate bankruptcy. For example, Serrano-Cinca (1996) has studied the applicability of self-organizing feature maps (SOFM) in the financial sector. Yang and Honavar (1997) have applied the probabilistic neural network (PNN) in bankruptcy prediction, and compared it with the backpropagation neural network and DA. Their results showed that PNN led to better prediction performance.

The bankruptcy prediction ability and excellent predictive accuracy of ANN in various areas has been proven by many studies. However, ANN has been criticized for some critical weaknesses, such as its potential to overfit training data and its lack of ability to explain results. Overcoming the risk of overfitting is significant, because bankruptcy prediction often requires huge data sets for generalizing results.

As a result, recent studies on bankruptcy prediction have popularly adopted the support vector machine (SVM) as an alternative to ANN. SVM is known as a technique that not only produces accurate prediction results but also enables training with small samples. It is also free from the risk of overfitting. As a result, several studies published in the mid-2000s, such as those of Park et al. (2005), Shin et al. (2005), and Min

and Lee (2005), have adopted SVM for bankruptcy prediction, and report that SVM outperforms all other comparative algorithms.

Recent studies on bankruptcy prediction have also researched the application of SVM and its modified algorithms. For example, Shin and Hong (2011) have applied AdaBoost algorithm-based SVM in a bankruptcy prediction model for IT companies in Korea. In their study, the authors performed multi-class credit ratings of companies by making a normal distribution shape for posterior bankruptcy probabilities from the loss functions extracted from SVMs. The results showed that their proposed method could minimize misclassification costs.

Zhou et al. (2014) have proposed the use of least-square SVM models for bankruptcy prediction, through the use of a new approach based on direct search and features that rank technology to optimize feature selection and parameter settings.

Choi and Ahn (2015) have suggested a novel model for improved bankruptcy prediction that converges three techniques: SVM, π-fuzzy function, and genetic algorithm (GA). Their bankruptcy prediction model is essentially based on SVM, but also incorporates fuzzy theory to extend the dimensions of the input variables. In addition, the authors propose the adoption of GA to optimize controlling parameters and feature subset selection.

Table 1 summarizes existing literature and the methods used to predict corporate bankruptcy. Overall, it shows that bankruptcy prediction studies have evolved with the development of classification techniques. Although better data preprocessing procedure may also lead to better prediction performance, studies on proper data preprocessing for bankruptcy prediction have thus far seldom been published.

〈Table 1〉 Prior research on the prediction of corporate bankruptcies

Reference	Proposed Classifier	Benchmark
Altman(1968)	DA	-
Deakin(1974)	DA	-
Ohlson(1980)	LR, PR	-
Odom and Sharda(1990)	BPN	DA
Tam and Kiang(1992)	BPN	DA, LR, k-NN, DT
Serrano-Cinca(1996)	SOFM	-
Yang and Honavar(1997)	PNN	BPN, DA
Park et al.(2005)	SVM	BPN, LR, DA
Shin et al.(2005)	SVM	BPN
Min and Lee(2005)	SVM	BPN, DA, LR
Tai and Shin(2009)	GA	-
Shin and Hong(2011)	SVM	LR, BPN
Zhou et al.(2014)	SVM	-
Choi and Ahn(2015)	SVM	LR, DA, DT, BPN, k-NN

(DA, Discriminant Analysis; LR, Logistic Regression; PR, Probit Regression; BPN: Backpropagation Network; SOFM, Self Organizing Feature Maps; PNN, Probabilistic Neural Network; SVM, Support Vector Machine; k-NN: k-Nearest Neighbor; DT: Decision Tree)

2.2. Imbalanced Data Handling

The number of insolvent companies in real

world forms an imbalance on the data much less than the number of healthy companies. The ratio of bankrupt companies to non-bankrupt companies is generally as low as 5 to 95 or below. Nevertheless, only a few conventional studies on bankruptcy prediction have taken into account the imbalance problem(Zhou, 2013).

When we use an unbalanced data to develop bankruptcy prediction models without any preprocessing, classification algorithms predict most cases of the minority class(i.e. bankrupt cases) to those of majority class(i.e. non-bankrupt cases). That is, the classification algorithms are easy to be overwhelmed by the majority class, and to ignore the minority class(Sundarkumar and Ravi, 2015). Consequently, it is required to apply sampling to the original data set before building classification models.

To tackle the imbalance problem of real-world datasets for bankruptcy prediction, most prior studies have been used dataset with paired samples, in which the number of non-bankrupt companies is the same as that of the bankrupt companies. However, some recent studies on imbalanced data handling have shown that the selection of proper sampling approach may lead to better prediction performance(Garcia et al., 2012; Zhou, 2013).

Generally, there are two types of sampling approaches to adjust the class distribution of a dataset: under-sampling and over-sampling. Under-sampling is a method of decreasing the number of majority class data points by eliminating majority class data points currently in the training set(Liu et al., 2007). Under-sampling has the advantage of a short training time(Liu et al, 2009). However, it has the risk of ignoring potentially useful data.

Several interesting works on the use of under-sampling have recently been published. For example, Jindaluang et al.(2014) applied a clustering algorithm that is performance guaranteed, named k-centers algorithm which clusters the data in many proportions, and then combines them with all the data in the minority class as a training set. They compared their approach with k-means on five data sets from UCI with two classifiers: 5-nearest neighbors and C4.5 decision tree, and found that the proposed sampling approach outperformed the comparison approach.

Wang and Shi(2014) proposed a density-weighted under-sampling method for SVM on imbalanced data. In their approach, density of points in majority group were first calculated. Then, the region growing method was employed to separate majority data into clusters and the centers of clusters are considered as the representatives of majority group. The seeds of region growing were selected randomly in proportion with data density. This choice of seed selection tends to take boundary points as the survivors. The sampled data were then built by putting cluster and minority group together. Experiments showed that their proposed approach yielded better prediction performance than random under-sampling method and CNN(Condensed Nearest Neighbor) – a conventional under-sampling approach proposed by

Hart(1968).

Ng et al.(2015) also proposed a novel under-sampling approach called diversified sensitivity-based under-sampling. In their approach, the samples of the majority class were clustered to capture the distribution information and enhance the diversity of the resampling. The proposed method showed a good generalization capability for 14 UCI datasets.

Sundarkumar and Ravi(2015) proposed a novel hybrid under-sampling strategy that employs k-Reverse Nearest Neighborhood and SVM in tandem. To validate the effectiveness of their approach, the authors applied it to the prediction of insurance fraud and customer churn. Empirical results showed that decision tree(J48) and SVM with their proposed approach yielded higher sensitivity.

Contrast to under-sampling, over-sampling adjusts the class distribution of a data set by increasing the number of minority class data points by sampling with replacement. It is known to be effective when the number of the minority class is extremely low. However, it has the risk of falling into overfitting. The most popular over-sampling strategy is the Synthetic Minority Over-sampling Technique(SMOTE) proposed by Chawla et al.(2002). It is designed to generate synthetic examples by operating in 'feature space' rather than 'data space' to overcome the overfitting (Chawla et al., 2002).

Other than SMOTE, several researchers have proposed novel over-sampling approaches. For example, Liu et al (2007) introduced the generative over-sampling algorithm, a resampling algorithm that creates artificial data points form a probability distribution learned from the minority class. Empirically, the results showed that generative over-sampling works well for a range of text classification datasets using linear SVMs.

Anitha and Santhi(2015) also introduced the new method namely Minority Over-sampling. This method is technique for imbalanced dataset learning using agglomerative clustering. It is designed to find the hard-to-learn informative minority samples first, and then assign weights to the minority class samples based on the Euclidean distance which are nearer to the majority class samples.

Lee and Kwon(2013) proposed a hybrid model of SVM, ANN and decision tree combined with the over-sampling approach from original imbalanced data set in order to improve specificity while maintaining sensitivity. To evaluate the performance of the proposed model, the study applied it to churn prediction, and it found that the proposed hybrid SVM model with over-sampling approach improved the specificity as expected.

Decision making on selecting under-sampling or over-sampling for an imbalanced dataset may be affected by the condition of the dataset. According to Zhou(2013) which compared the performances of under-sampling approaches with ones of over-sampling approaches under various conditions, under-sampling is better than over-sampling when there are hundreds observations of the minority class in the dataset.

2.3. Imbalanced Data Handling in Bankruptcy Prediction

As mentioned above, few existing studies have addressed the issue of imbalance in bankruptcy prediction. Kotsiantis et al. (2007) suggest that a systematic study be carried out on the various methodologies that have been used to handle the problem of imbalanced datasets. Such a study could create ensembles of classifiers by distributing a training set so that balance is reached in each of the resulting training samples. Through this methodology, one could improve the identification of difficult small classes (e.g., bankrupt firms) in a predictive analysis, while maintaining the classification ability of other classes (e.g., non-bankrupt firms) at an acceptable level.

Zhou (2013) has investigated the effects of various sampling methods on the performance of quantitative bankruptcy prediction models in real highly-imbalanced datasets. To compare models' performances, tests on random paired sample sets and real imbalanced sample sets were conducted by the author and compared to each other. The results showed that the proper sampling method for developing prediction models mainly depends on the number of bankruptcies in the training sample set.

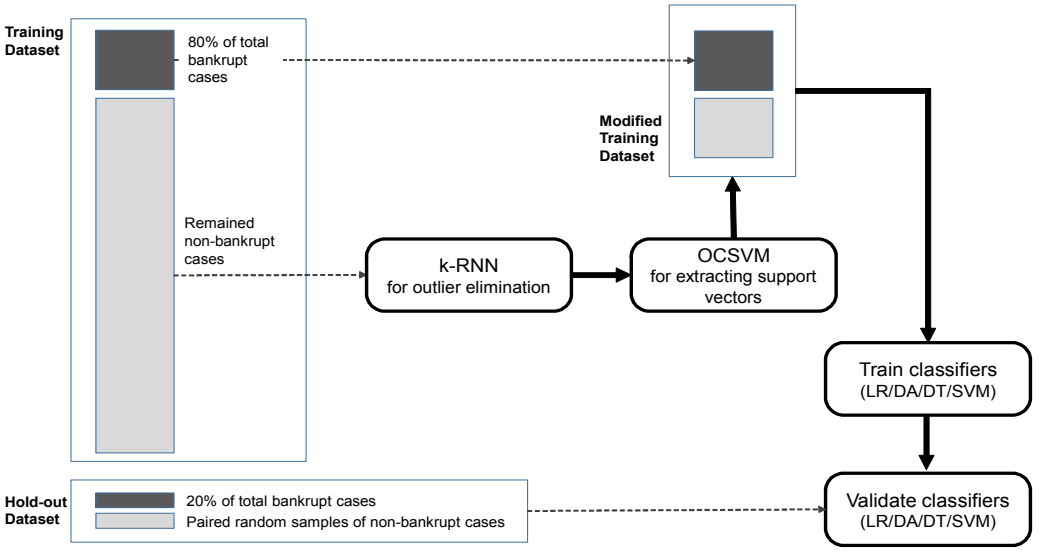
Kim et al. (2015) have proposed using the geometric mean based boosting algorithm (GMBoost) to resolve data imbalance problems. They applied GMBoost to a bankruptcy prediction task in order to evaluate its performance. The results showed that, in a comparative analysis involving AdaBoost and cost-sensitive boosting, GMBoost had the advantage of greater prediction power and more robust learning capabilities when it was used with imbalanced data.

As discussed, only a few researchers have proposed using sampling approaches to address imbalanced data problems in bankruptcy prediction. However, as those studies were published recently, we can safely say that this research topic is one that is currently gaining a great amount of attention.

3. Proposed Approach

In this study, we propose a novel hybrid under-sampling approach to bankruptcy prediction. A recent study by Zhou (2013) has indicated that under-sampling is more effective than over-sampling for bankruptcy prediction, since the numbers of bankruptcy cases are in the hundreds or above in most cases. This is why we have chosen to use under-sampling approaches for bankruptcy prediction.

Of the various techniques available for under-sampling, we propose a combination of k-RNN and OCSVM, as shown in Figure 1. k-RNN is expected to eliminate outliers effectively, while OCSVM identifies informative training samples from the data of the majority class. Sundarkumar and Ravi (2015) have shown how a combination of k-RNN and OCSVM can lead to better accuracy in fraud detection and



〈Figure 1〉 Procedure of the proposed approach

customer churn prediction. Our study proposes this under-sampling approach as a means of improving bankruptcy prediction.

In order to apply our proposed approach, the hold-out dataset must first be extracted, comprising 20% of the original unbalanced dataset and using stratified random sampling. Then, the remaining 80% of the dataset becomes subject to our proposed approach. In our approach, k-RNN is applied first to eliminate outliers from the majority class. After that, OCSVM is used to extract the support vectors of the majority class. Since these support vectors form the boundaries of the majority class, they can be informative in representing the properties of the majority class as a whole. In addition, by applying k-RNN and OCSVM, we can eliminate a huge number of useless samples from the majority class, which makes it possible to indirectly under-sample. The samples selected through k-RNN and OCSVM can then be merged with all the remaining samples of the minority class, thereby producing a modified and balanced dataset. Figure 1 depicts the overall procedure involved in our proposed approach.

The detailed explanation on the core sampling algorithms, that is, k-RNN and OCSVM, is as follows:

3.1. k-Reverse Nearest Neighbor

k-RNN, as proposed by Soujanya et al. (2006), is a technique for eliminating outliers by using the concept of the k-NN(k-Nearest Neighbor) algorithm. It operates as follows. First, let X be a dataset with n samples, $X = \{x_1, x_2, x_3, \cdots, x_i, x_j, \cdots, x_n\}$. Then, the set of k-NN($x_i$) becomes $\{x_j | d_{ij} \leq$ kth nearest

distance of $x_i\}$ where d_{ij} is the distance between two samples x_i and x_j. In this situation, the set of k-RNN(x_i) is defined as $\{x_j|x_i \in$ k-NN(x_j) $\}$. This means that a point x_j belongs to k-RNN(x_i) if and only if x_i belongs to k-NN(x_j). In k-NN, each sample in the dataset always has at least k nearest neighbors for a given k. However, k-RNN does not ensure the existence of the reverse neighbors. When a sample in the dataset is very far from other samples, its set of k-RNN may have no elements. On the other hand, if a point x_i has many k-RNNs, x_i is regarded as having a denser neighborhood, which implies a lower probability of being an outlier (Kumar et al., 2011). This is why k-RNN can be used to identify outliers from a dataset.

In a general form, an outlier point is defined one that has less than K number of k-RNNs. That is, the set of outliers of *X* becomes $\{x_i|$ |k-NN(x_i)| $\leq$ K $\}$. Thus, in order to use k-RNN for identifying outliers, researchers should determine the appropriate k(the number of neighbors in k-NN) and K(the minimum number of reverse neighbors that is required to not be selected as outliers in k-RNN). In general, the higher the difference is between k and K, the higher the probability of a sample is an outlier. In addition, when k has too small of a value or too large of a value, it may hinder k-RNN from identity outliers (Sundarkumar and Ravi, 2015). Thus, the optimal values of k and K must be determined by trial-and-error.

3.2. One-class Support Vector Machine

The conventional SVM proposed by Vapnik (1998) performs classification by mapping input vectors onto an *N*-dimensional feature space, and constructing a linear model called "the optimal hyperplane," which implements nonlinear class boundaries in the original space. Technically, the optimal hyperplane separates out the training examples with the maximum distance from the hyperplane to the closest training data samples. Those training examples that are closest to the optimal hyperplane are called support vectors. Support vectors represent the boundaries of each class, and so can be understood as informative training samples.

OCSVM is quite different from conventional SVM. In contrast with SVM, it deals with training data that have only one class. The goal of OCSVM is to describe the data from a single class by using support vectors. From the technical perspective, it builds a boundary that separates a class from the rest of the feature space (Sundarkumar and Ravi, 2015; Tax and Duin, 2004). Since our goal is to select the appropriate samples from the samples within the majority class (i.e., a single class), it is more appropriate to use OCSVM than SVM. The detailed process of formulating and solving OCSVM has been discussed in Tax and Duin (2004). For more details, please refer to that source, as listed in our references.

4. Empirical Validation

4.1. Research data

To test the effectiveness of our proposed approach, we carried out an empirical analysis using H Bank's bankruptcy data for non-externally audited companies in Korea. In Korea, non-externally audited companies are those whose total assets range from 1 billion KRW to 7 billion KRW. Companies whose total assets exceed 7 billion KRW must undergo an external audit. Non-externally audited companies may often have problems with the transparency of accounting information. Consequently, it is known that bankruptcy predictions on non-externally audited companies are more difficult to work with than those by externally audited companies.

The dataset for our empirical test comprised 163 financial ratios and one class variable. A value of 0 (non-bankrupt) or 1 (insolvent) was assigned to the class variable. In this case, insolvency refers to bankruptcy or a condition in which payments are more than three months overdue. In order to build an industry-specific bankruptcy prediction model, we limited the bankruptcy dataset to "heavy industry" companies. The total number of samples in our dataset was 8,326. Among them, there were 774 (9.3%) insolvent cases and 7,552 (90.7%) non-bankrupt cases.

Since there were too many independent variables in our dataset, we finally selected nine (9) financial ratios by using an independent sample t-test, engaging in variable selection through logistic regression, and considering expert opinion. The selected independent variables included (1) coefficient of sales volatility, (2) financial costs to sales, (3) accumulated earnings to total assets, (4) total borrowings and bonds payable to total assets, (5) cash to current liabilities, (6) volatility of working capital to sales, (7) borrowings to EBITDA, (8) trade payable turnover period, and (9) cash flow to debts.

The missing values were replaced by the arithmetic means of the corresponding variables. Min-max normalization was also applied, to ensure that the larger-value input features did not overwhelm the smaller-value input features. To limit the effects of outliers, we also applied Winsorization, which sets all outliers to a specified percentile of the data. Considering the distribution of our dataset, we used 90% Winsorization. As such, all data below the 5th percentile was set to the 5th percentile, and all data above the 95th percentile was set to the 95th percentile.

4.2. Data Preprocessing

In accordance with the procedure of our approach, as depicted in Figure 1, we partitioned our dataset into two subsets (training and hold-out). To do so, we first extracted the hold-out dataset (154 bankruptcy cases and 154 non-bankruptcy cases) by using stratified random sampling. The hold-out dataset was set aside to validate the usefulness of our approach. The remaining dataset (620 bankruptcy cases and 7,398 non-bankruptcy cases) was used as the training

dataset. The 7,398 non-bankruptcy cases in the training dataset thus became the inputs of our proposed under-sampling approach.

As the first step in our approach, we applied k-RNN. As explained in Section 3.1, the values of k and K must be set arbitrarily in order to apply k-RNN. Through a fine-tuning procedure, we finally chose 5 as k and 8 as K. In general, a sample is regarded as an outlier if it has fewer reverse neighbors (K) than k (Sundarkumar and Ravi, 2015). Our settings (5 for k and 8 for K) thus indicate that we applied a very tight standard for identifying outliers. Accordingly, 6,091 cases (82.33%) were identified as outliers and removed.

We then employed OCSVM to extract support vectors from residual cases (1,307 cases). As a result, we extracted 655 support vectors. Because the total number of bankruptcy samples in the training dataset was 620, the class distribution ratio of majority class to minority class in the modified training dataset was almost balanced.

4.3. Experimental design

To implement k-RNN, we employed our own program, written in Microsoft VBA for Excel 2013. For OCSVM, we used an open-source software called LIBSVM v3.20 (Chang and Lin, 2011) and a post-processing software that extracts support vectors and identifies them in the training dataset, within a Microsoft Excel environment.

With regard to the classification methods, we used four different classifiers: logistic regression (LR), discriminant analysis (DA), classification and regression trees (CART), and support vector machines (SVM). LR, DA, and CART were employed by using IBM SPSS for Windows Ver 20. In the case of SVM, prediction performance is affected by kernel function and its parameters. Since there are few guidelines for determining kernel function and its parameters in SVM, we varied them to search for the optimal values, and applied all three types of kernel function: the linear function, the polynomial function, and the Gaussian radial basis function (RBF). We set the values with σ^2 as 1, 25, 50, 75, and 100, and with C as 10, 33, 55, 78, and 100 for the Gaussian RBF. For the polynomial function and linear kernel function, we set the values of C as 10, 33, 55, 78, and 100. For d (the degree of polynomial function), we used a range from 1 to 5 (Ahn and Kim, 2009). We also used LIBSVM v3.20 to implement support vector classification (Chang and Lin, 2011).

To validate the competence of our proposed approach, we compared the performances of classifiers under our proposed under-sampling approach and a simple random under-sampling approach. As the criteria for evaluating prediction performance, we used both overall hit ratio (%) and false negative rate (%), which represent the misclassification costs of bankruptcy prediction. These two measures can be calculated as follows:

Overall hit ratio (%)
= (TP + TN) / (TP + TN + FP + FN)
False negative rate (%) = FN / (FN + TP)

where TP stands for true positive, TN stands for true negative, FP stands for false positive, and FN stands for false negative (refer to Table 2).

〈Table 2〉 Confusion Matrix

		Actual	
		Bankrupt	Non-bankrupt
Prediction	Bankrupt	True positive	False positive
	Non-bankrupt	False negative	True negative

4.4. Experimental results

<Table 3> presents the overall hit ratios of the models under our proposed under-sampling approach, while Table 4 presents those under the simple random under-sampling approach. By comparing these tables, we can see that our hybrid under-sampling approach slightly enhances the prediction accuracies for LR, DA, and CART.

〈Table 3〉 Overall hit ratios of the proposed under-sampling approach

Classifier	Training	Hold-out	Settings
LR	66.21%	67.53%	
DA	65.89%	67.21%	
CART	68.47%	64.29%	Max. depth = 5, Gini index
SVM	67.82%	69.48%	Polynomial kernel, C=100, d=2

〈Table 4〉 Overall hit ratios of the simple random under-sampling approach

Classifier	Training	Hold-out	Settings
LR	64.55%	67.86%	
DA	64.94%	67.53%	
CART	67.92%	65.26%	Max. depth = 5, Gini index
SVM	65.88%	66.56%	Linear Kernel, C=33

<Table 5> presents the false negative rates for the models, for both our hybrid under-sampling and simple random under-sampling approaches. As shown in the table, our proposed approach has been proven to reduce the misclassification cost of false negative errors, regardless of the classifier type. In particular, our approach helped to dramatically reduce the false negative error rate for SVM, although it failed to improve the overall hit ratio of SVM. Considering that the misclassification of bankrupt companies as non-bankrupt ones causes huge costs for financial institutions, we can conclude that our approach is particularly valuable for bankruptcy prediction.

〈Table 5〉 Comparison of false negative rates

Classifier	Simple Random Under-sampling	Our Hybrid Approach (k-RNN + OCSVM)
LR	29.87%	18.18%
DA	31.17%	16.23%
CART	41.56%	38.96%
SVM	23.37%	11.04%

5. Conclusion

In this study, we have proposed a new hybrid under-sampling approach that combines k-RNN and OCSVM for better bankruptcy prediction. This approach was first proposed by Sundarkumar and Ravi (2015), but has never been applied for bankruptcy prediction. With this in mind, we validated the effectiveness of the approach for bankruptcy prediction by applying it

to real-world bankruptcy prediction samples from Korea.

Our dataset was very imbalanced, because only 9.3% of the cases in the overall dataset involved bankruptcy. However, since the total size of the dataset was quite large, the sample size of this minor class included more than 700 items. Given this, we determined that under-sampling would be more appropriate than over-sampling for our bankruptcy prediction dataset, and thus applied the hybrid approach of k-RNN and OCSVM to the dataset.

Our empirical results showed that our approach improved the overall accuracy of LR, DA, and CART. They also showed that it significantly reduced the false negative rate of LR, DA, CART, and SVM, as compared to simple random under-sampling. Since our proposed approach has been proven to significantly reduce the misclassification costs of bankruptcy prediction models, it can be used effectively in practice.

From the academic perspective, only a few studies have dealt with handling data imbalance problems in bankruptcy prediction. As such, our study makes an important contribution, by emphasizing the importance of proper sampling for effective bankruptcy prediction.

In particular, our proposed approach was empirically proven that it leaded to reduce false negative errors. Thus, from the practical perspective, our approach is expected to help the financial institutions to avoid their critical mistakes such as making a loan to inappropriate debtors.

On the other hand, this study has certain limitations. First, our results were improved by our approach, but not at a significant level. Our lack of data may have been a critical factor in this result. Therefore, in future studies, it will be necessary to verify the effectiveness of our proposed approach by using more extensive data.

Second, a mechanism for optimizing the parameters of our proposed approach is required. In particular, the values of k and K in k-RNN seriously affected the performance of our proposed approach, but we have thus far had to determine these values with our heuristic. Optimization through the use of a genetic algorithm (GA) may be one alternative for overcoming this challenge. With this in mind, studies need to investigate how to optimize the parameters that need to be followed in the future.

Third, a test of the applicability of other sampling approaches is needed. In general, over-sampling is known to be a more effective approach to prediction than under-sampling, because it is free of information loss. Thus, there may be a need to apply some over-sampling approaches (e.g., SMOTE) to bankruptcy prediction, and to compare their performances with those of under-sampling approaches.

References

Ahn, H., and K.-j. Kim., "Corporate Bond Rating using Various Multiclass Support Vector Machines." *Asia Pacific Journal of Information Systems*, Vol.19, No.2(2009), 157~178.

Altman, E. I., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, Vol.23, No.4(1968), 589～609.

Anitha, R., and S. Santhi., "Minority Oversampling Technique for Imbalanced Dataset Learning Using Agglomerative Clustering," *International Journal of Emerging Technology and Innovative Engineering,* Vol. 1, No.3(2015), 137～142.

Bellovary, J. L., D. E. Giacomino, and M. D. Aker, "A Review of Bankru,ptcy Prediction Studies: 1930 to Present," *Journal of Financial Education,* Vol.33, No.4(2007), 1～43.

Chang, C. -C. and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* Vol.2, No.3(2011), 1～27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chawla, N. V., K. W. Bowyer, and L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* Vol.16(2002), 321～547.

Choi, S. Y., and H. Ahn, "Optimized Bankruptcy Prediction through Combining SVM with Fuzzy Theory," *Journal of Digital Convergence*, Vol.13, No.3(2015), 155～165.

Deakin, E., "A Discriminant Analysis of Predictors of Business Failure," *Journal of Accounting*, Vol.10, No.1(1974), 167～179.

Garcia, V., J. S. Sanchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, Vol. 25(2012), 13～21.

Hart, P. E., "The Condensed Nearest Neighbor Rule," *IEEE Transactions on Information Theory*, Vol. 18, (1968), 515～516.

Jindaluang, W., V. Chouvatut, and S. Kantabutra, "Under-sampling by algorithm with performance guaranteed for class-imbalance problem," *Computer Science and Engineering Conference (ICSEC),* (2014), 215～221.

Kim, M. J., D. K. Kang, and H.B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Expert Systems with Applications*, Vol.42, No.3 (2015), 1074～1082.

Kim, S., C. S. Park, and S. M. Jeon, "Default Decisions of FIs and Endogeneity Problems in Default Prediction," *Journal of Business Research,* Vol.26, No.1(2011), 99～132.

Kotsiantis, S., D. Tzelepis, E. Koumanakos, and V. Tampakas, "Selective costing voting for bankruptcy prediction," *International Journal of Knowledge-based and Intelligent Engineering Systems*, Vol.11(2007), 115～127.

Kumar, P. and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques-A review," *European Journal of Operational Research*, Vol.180, No.1(2007), 1～28.

Kumar, P., P. R. Krishna, and S. B. Raju, *Pattern Discovery Using Sequence Data Mining: Applications and Studies: Applications and Studies*, IGI Global, Hershey, Pennsylvania, 2011.

Lee, J. S. and J. G. Kwon, "A Hybrid SVM Classifier for Imbalanced Data Sets," *Journal of Intelligence and Information Systems,* Vol.19, No.2(2013), 125～140.

Liu, A., J. Ghosh, and C. E. Martin, "Generative Oversampling for Mining Imbalanced Datasets," *Proceedings of the 2007 International Conference on Data Mining,* (2007), 66～72.

Liu, X. Y., J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* Vol.39 No. 2(2009), 539～550.

Min, J. H. and Y.-C. Lee, "Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters," *Expert Systems with Applications*, Vol.28, No.4(2005), 603～614.

Ng, W. W., J. Hu, D. S. Yeung, S. Yin, and F. Roli, F, "Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems," *IEEE Transactions on Cybernetics,* (2015), Forthcoming.

Odom, M. D., and R. Sharda, "A Neural Network Model For Bankruptcy Prediction," *Proceedings of the International Joint Conference on Neural networks,* Vol.2(1990), 163～168.

Ohlson, J. A., "Financial Ratios and the Probabilistic Prediction of Bankruptcy," *Journal of Accounting Research,* Vol.18, No.1(1980), 109～131.

Park, J.-m., K.-j. Kim, and I. Han, "Bankruptcy Prediction using Support Vector Machines," *Asia Pacific Journal of Information Systems*, Vol.15, No.2(2005), 51～63.

Serrano-Cinsa, C., "Self organizing neural networks for financial diagnosis," *Decision Support Systems,* Vol.17, No.3(1996), 227～238.

Shin, K.-S., T. S. Lee, and H.-j. Kim, "An application of support vector machines in bankruptcy prediction model," *Expert Systems with Applications,* Vol.28, No.1(2005), 127～135.

Shin, T. and T. Hong, "Corporate Credit Rating based on Bankruptcy Probability Using AdaBoost Algorithm-based Support Vector Machine," *Journal of Intelligence and Information Systems,* Vol.17, No.3(2011), 25～41.

Soujanya, V., R. V. Satyanarayana, and K. Kamalakar, "A Simple Yet Effective Data Clustering Algorithm," *Proceedings of the Sixth International Conference on Data Mining(ICDM'06), Hong Kong,* (2006), 1108～1112.

Sundarkumar, G. G. and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *Engineering Applications of Artificial Intelligence,* Vol.37, (2015), 368～377.

Tai, Q.-y., and K.-s. Shin, "GA-based Normalization Approach in Backpropagation Neural Network for Bankruptcy Prediction Modeling," *Journal of Intelligence and Information Systems,* Vol.15, No.3(2009), 1～14.

Tam, K. Y. and M. Y. Kiang, " Managerial Applications of Neural Networks : The Case of Bank Failure Predictions," *Management science*, Vol.38, No.7(1992), 926～947.

Tax, D. M. J., and R. P. W. Duin, "Support Vector Data Description," *Machine Learning*, Vol. 54, No.1(2004), 45～66.

Vapnik, V. N., *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

Wang, D., and M. Shi, "Density Weighted Region

Growing Method for Imbalanced Data SVM Classification in Under-sampling Approaches," *Journal of Information & Computational Science,* Vol.11, No.18(2014), 6673～6680.

Yang, J., and V. Honavar, "Feature Subset Selection Using a Genetic Algorithm," *Computer Science Technical Reports,* (1997), Paper 156.

Zhou, L., "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling method," *Knowledge-Based Systems,* Vol.41(2013), 16～25.

Zhou, L., K. K. Lai, and J. Yen, "Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation," *International Journal of Systems Science*, Vol.45, No.3(2014), 241～253.

국문요약

부도예측 개선을 위한 하이브리드 언더샘플링 접근법

김태훈* · 안현철**

부도는 막대한 사회적, 경제적 손실을 야기할 수 있으므로, 미리 부도여부를 정확하게 예측하여 선제 대응하는 것은 경영분야에서 대단히 중요한 의사결정문제 중 하나이다. 이에 지능정보시스템 분야에서도 그간 기업의 재무 데이터에 기반해 부도예측을 개선하기 위한 노력을 기울여왔는데, 안타깝게도 기존의 연구들은 대부분 분류모형의 성능 개선을 통해 예측 정확도를 개선하는 것에만 주로 초점을 맞추어 다른 요소들을 충분히 고려하지 못했다는 한계가 있다. 이러한 배경에서 본 연구는 부도예측 모형의 정확도를 개선하기 위한 방편으로 새로운 데이터 전처리 방법, 그 중에서도 효과적인 표본추출 방법을 제안하고자 한다. 일반적으로 부도예측을 위해 사용되는 데이터들은 극심한 데이터 불균형 문제에 노출되어 있는데, 본 연구에서는 k-reverse nearest neighbor(k-RNN)와 one-class support vector machine(OCSVM) 방법을 결합한 하이브리드 언더샘플링(hybrid under-sampling) 접근법을 통해 이같은 데이터 불균형 문제를 해결하고자 하였다. 본 연구에서 제안한 접근법에서 k-RNN은 이상치를 효과적으로 제거할 수 있으며, OCSVM은 다수를 구성하는 등급의 데이터로부터 정보량이 풍부한 표본만 효과적으로 선택할 수 있는 수단으로 활용될 수 있다. 제안된 기법의 성능을 검증하기 위해, 본 연구에서는 국내 한 은행의 비외감기업 부도예측모형 구축에 제안 기법을 적용해 본 뒤, 일반적으로 많이 사용되는 랜덤샘플링(random sampling)과 제안 기법의 성능을 비교해 보았다. 그 결과, 로지스틱 회귀분석, 판별분석, 의사결정나무, SVM 등 대다수의 분류모형에 있어 분류 정확도가 개선됨을 확인할 수 있었으며, 모든 분류모형에 있어 부정 오류, 즉 부실기업을 정상으로 예측하는 오류율이 크게 감소함을 확인할 수 있었다.

주제어 : 부도예측, 언더샘플링, k-Reverse Nearest Neighbor, One-class Support Vector Machine, 분류

논문접수일 : 2015년 5월 20일 논문수정일 : 2015년 6월 15일 게재확정일 : 2015년 6월 16일
투고유형 : 영문급행 교신저자 : 안현철

* 국민대학교 비즈니스IT전문대학원 석사과정
** 교신저자: 안현철
국민대학교 비즈니스IT전문대학원 부교수
서울특별시 성북구 정릉로 77, 136-702
Tel: (02)910-4577, Fax: (02)910-5209, E-mail: hcahn@kookmin.ac.kr

저 자 소 개

김 태 훈
국민대학교 경영정보학부에서 학사, 비즈니스IT전문대학원에서 비즈니스IT전공으로 석사과정에 재학 중이다. 주요 관심분야는 빅데이터 분석이다.

안 현 철
현재 국민대학교 경영대학 경영정보학부 부교수로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심분야는 금융 및 고객관계관리 분야의 인공지능 응용, 정보시스템 수용과 관련한 행동 모형 등이다.