IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 3, 2019, accepted January 27, 2019, date of publication January 30, 2019, date of current version February 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896474

Application of Instance-Based Entropy Fuzzy Support Vector Machine in Peer-To-Peer Lending Investment Decision

POONGJIN CHO[1], WOOJIN CHANG[1], AND JAE WOOK SONG[2]
[1]Department of Industrial Engineering, Institute for Industrial Systems Innovation, Seoul National University, Seoul 08826, South Korea
[2]Department of Data Science, Sejong University, Seoul 05006, South Korea

Corresponding author: Jae Wook Song (jaewook.song@sejong.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant through the Ministry of Science and ICT, under Grant 2018R1C1B5043835.

ABSTRACT Loan status prediction is an effective tool for investment decisions in peer-to-peer (P2P) lending market. In P2P lending market, most borrowers fulfill the repayment plan; however, some of them fail to pay back their loans. Therefore, an imbalanced classification method can be utilized to discriminate such default borrowers. In this context, the aim of this paper is to propose an investment decision model in P2P lending market which consists of fully paid loans classified via the instance-based entropy fuzzy support vector machine (IEFSVM). IEFSVM is a modified version of the existing entropy fuzzy support vector machine (EFSVM) in terms of an instance-based scheme. IEFSVM can reflect the pattern of nearest neighbors entropy with respect to the change of its size instead of fixing it in unified neighborhood size. Therefore, IEFSVM allows the class change of nearest neighbors in the determination of fuzzy membership. Applying the model to the lending club dataset, we determine loans that are predicted to be fully paid. Then, we also provide a multiple regression model to generate an investment portfolio based on non-default loans that are predicted to yield high returns. Throughout the experiment, the empirical results reveal that IEFSVM outperforms not only EFSVM but also the six other state-of-the-art classifiers including the cost-sensitive adaptive boosting, cost-sensitive random forest, EasyEnsemble, random undersampling boosting, weighted extreme learning machine, and cost-sensitive extreme gradient boosting in terms of loan status classification. Also, the investment performance of the multiple regression model using IEFSVM is higher and more robust than that of two other benchmarks. In this regard, we conclude that the proposed investment model is a decent and practical approach to support decisions in the P2P lending market.

INDEX TERMS Entropy, support vector machines, financial management, decision support systems, peer-to-peer lending.

I. INTRODUCTION

Peer-to-peer (P2P) lending, one of the most well-known financial technology, is a service that connects individual investors with loan borrowers through online platform. When borrowers apply for a loan in the platform, many investors lend money and receive interest for a fixed period of time. P2P market is cheaper than credit card loans with higher investment return and lower operating cost than commercial banks. P2P lending platform evaluates each loan application based on its own algorithm to immediately decide whether to approve the loan [1]. Since investors can easily identify all loan applications, they can invest online based on the information without any difficulty. In addition, entry barriers for borrowers are significantly reduced as investors are willing to loan in small amounts [2]. Furthermore, factors such as deregulations in financial sector, technological development, and need for revenue models have continuously contributed to increase the size of the P2P lending market.

The aim of this research is to provide an investment decision model in P2P lending market based on an imbalanced classification algorithm and multiple linear regression for loan evaluation. When evaluating loan applications, borrowers provide their personal and financial information

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

into the P2P platform such as annual income and interest rate. Then, the P2P platform not only decides whether to approve the loan but also imposes the rating of borrower's risk. For instance, the Lending Club, the largest P2P lending platform, grades each loan application from A to G, and in detail from A1 to A5, . . . , G1 to G5. Obviously, the average return and the risk of grading are different for each grade. The closer to A-grade, the more trustworthy the borrowers are, so they can borrow a large amount at a low interest rate. On the other hand, the creditworthiness of borrowers is very low when a rating is close to G-grade, so they borrow money at a high interest rate with limited amount at a time. From an investor's point of view, the average interest rate of A-grade loan is low at 7.52%, which yields the investment return at 7.12%. Although the interest is low, most loans in A-grade are fully paid with the low average risk at 13.25% given that these loans are highly creditworthy. In contrast, G-grade loans are expected to yield a high investment return with the high average interest rate at 23.75%. However, there are many default loans which yield the average investment return to be reduced to 11.59%. In addition, the average risk is 37.59%, which is much higher than that of A-grade.

Such grading system enables a portfolio composition based on the risk aversion of investors. Note that the statistics of the Lending club data state that the fully paid loans are 87.4%, whereas the default loans are 12.6%. In this context, if we define the goal of this study as a classification problem to predict the final status of loans, the problem becomes an imbalanced classification given that the percentage of fully paid loans are much larger than that of default loans. Full returns on investment are only obtained from fully paid loans, whereas a principal can be lost in default ones. Therefore, investors should consider suck risks [3], and many related studies are underway on accurate grading system and investment strategy.

In general, the research topics of P2P lending market include loan evaluation, investment decision, and determinants of default. The loan evaluation problem, the most important topic among them, is essentially the same as the credit scoring problem in fixed income markets since it separates the minority data from the majority data. Therefore, many researchers have attempted to apply imbalanced classification methods. In case of credit scoring, various profit-based models have been proposed to optimize investment decisions for personal investors [4]–[6]. However, similar studies in P2P lending market are relatively small with limited understanding of the market. In recent studies, the loan status prediction problem has been researched in the perspective of newer techniques such as cost-sensitive version of extreme gradient boosting (XGBoost) [7], Bayesian hyper-parameter optimization of XGBoost [8], heterogeneous ensemble [9], soft information from descriptive text [10], and contrastive pessimistic likelihood estimation with gradient boosting [11], which could be considered as the state-of-the-art models. Hence, in this study, we compare the performance of our proposed model against the models above as benchmarks.

The first motivation of this research is to enhance the performance of the loan status prediction model. Specifically, we suggest to enhance the performance by integrating a concept from other research discipline. Fan *et al*. [12] attempted to develop the novel imbalanced classification method by applying the entropy to the fuzzy support vector machine. Entropy is known to possess explanatory power of data and to assess the degree of information's certainty [13]. In the field of data mining, developing new classifier with an entropy has been widely studied where Chen *et al*. [14] first proposed a method to measure uncertainty using entropy in a neighborhood system. Note that the neighborhood entropy can measure certainty in terms of classes for the prediction of classification problems. Also, the nearest neighbors entropy can be used to formulate the fuzzy membership of the fuzzy support vector machine [12]. However, a drawback of such method lies on the fact that it uses a uniform neighborhood size for all samples. Instead of considering the change of class elements based on the number of nearest neighbors, most of the existing literatures [15]–[19] allocated weights to the instances by focusing on developing proper function of neighborhood size. However, the unified neighborhood size is not suitable for some samples and may result in misclassification [20]. If the neighborhood size is too small, the important information can be lost. In contrast, if the neighborhood size is too large, it can cause the inclusion of outliers or the ignorance of a small amount of information in the nearest neighbors [21]. Hence, in order to develop a robust classifier by changing the neighborhood size, several studies [20], [22]–[25] combined information according to the number of nearest neighbors. Instead of tuning neighborhood size as a fixed value, it is reasonable to combine information in response to various neighborhood size to assign fuzzy membership of each data point. In this regard, we suggest to apply an instance-based entropy fuzzy support vector machine (IEFSVM), proposed in [26], in loan evaluation for P2P lending.

The second motivation of this paper is to improve the performance of the investment decision model in comarpsion to that of existing methods in P2P lending market. In previous literatures, Serrano-Cinca and Gutiérrez-Nieto [27] attempted to predict the expected profitability based on profit scoring through the internal rate of return (IRR). Based on the IRR, they constructed a portfolio investing in 100 loans that were expected to draw high IRR through regression model. Furthermore, Guo *et al*. [28] measured the credit risk of each loan by kernel regression. Since the model adjusted the weight of kernel with an instance-based model, it enabled the comparision of the return and risk of each loan. Then, they proposed a portfolio through optimization problem. Note that these studies evaluated the performances of their strategies based on the investment return and the Sharpe ratio [29]. On the other hand, some papers have developed a suitable classifier for imbalanced data and constructed a portfolio using the classifier itself. Xia *et al*. [7] developed a classifier

incorporating the cost-sensitive learning and XGBoost and proposed a portfolio allocation model with boundary constraints. In this milieu, we also propose an investment decision model, which constructs a portfolio by selecting the high-return loans based on IEFSVM and regression model. Note that the proposed model exhibits the higher investment return and the Sharpe ratio value than other models.

On the basis of two motivations, the contribution of this paper is combining a novel cost-sensitive loan status prediction model into an investment decision model specifically for the P2P lending market. The practical contribution of our model is providing a mechanism to predict fully paid loans based on IEFSVM. Furthermore, a simple regression model is used to rank loans that are predicted to yield high investment returns. Composing the high-ranked loans into the portfolio, our approach is expected to realize an investment decision with the high Sharpe ratio. The technical contribution of this research is improving the investment decision model proposed by Serrano-Cinca and Gutiérrez-Nieto [27] via discarding loans classified as default based on IEFSVM. This modification is also expected to promote creating a portfolio with the high Sharpe ratio.

The rest of this paper is organized as follows. Section 2 presents a brief review of the main related literatures; Section 3 describes IEFSVM approach, the component of IEFSVM for comparison, and the investment decision model; Section 4 discusses and analyzes the empirical results; and Section 5 concludes.

II. RELATED WORKS

Various models have been developed to enhance the profitability in P2P lending market. For example, Serrano-Cinca *et al*. [30] discovered the factors explaining defaults in P2P lending based on a hypotheses test and a survival analysis; Jiang *et al.* [10] extracted features from the descriptive loan text with latent Dirichlet allocation (LDA) model and predicted defaults in P2P loan using the soft information; and Zeng *et al*. [31] attempted to form a bipartite graph between loans and investors and built an iteration computation approach to develop an investment decision model. Since the P2P lending is similar to credit scoring, the literatures of the development of credit scoring are a decent precedent for the study in P2P lending [32]–[39]. Marques *et al*. [40] summarized the improvements of evolutionary algorithms to credit scoring. Among such studies, the most popular topic is the application of imbalanced classification. For instance, Sun *et al*. [41] integrated the synthetic minority over-sampling technique (SMOTE) with the Bagging ensemble and employed support vector machine (SVM) as the base classifier to predict financial distress based on imbalanced data sets. Also, Marqués *et al*. [42] demonstrated the suitability and the performance of several re-sampling techniques to the class imbalance problem in credit scoring.

The concept of nearest neighbor has been applied to deal with the class imbalance problem in data mining. For example, Chen *et al*. [43], [44] performed the ensemble learning with synthetized neighborhoods to handle the imbalanced data. Also, Ando [45] proposed to set the density of the nearest neighbor differently for each class with convex optimization technique. Furthermore, many researches applied the angle between a neighbor and mean of neighbors [13]–[15], the cosine sum of angles for all neighbors to weighted one-class support vector machine [15], sample reduction [16], and boundary detection [17] for support vector machine. In this regard, the concept of nearest neighbors entropy was developed to evaluate a measure of certainty to the belonging class of each nearest neighbor [14]. Especially, Fan *et al*. [12] proposed an entropy fuzzy support vector machine (EFSVM) to handle the class imbalance problem with the nearest neighbors entropy. However, EFSVM has the disadvantage of using a uniform neighborhood size for all data, which has always been a controversial problem when using nearest neighbors. Hence, some efforts have been made to address the drawbacks of the nearest neighbors. Pan *et al*. [20] predicted each class with the harmonic mean distance of nearest neighbors; Ertuğrul and Tağluk [22] measured the similarity and the dependency based on distance and angle between nearest neighbors; Zhu *et al*. [23] evaluated the distance of the fixed radius nearest neighbor with modified law of gravitation; and Zhang *et al*. [24] developed a dynamic local neighborhood concept taking positive-negative borders between each sample into account to alter the posterior probability of minority data. It is reasonable to combine the information with respect to the size of neighborhood [20], [22]–[25] instead of tuning neighborhood size as a fixed value. Since IEFSVM is committed to develop an appropriate combination of the nearest neighbors entropy without much dependence on the neighborhood size, it is expected to improve the performance of loan evaluation in P2P lending.

As mentioned above, in order to demonstrate the effectiveness of IEFSVM for loan evaluation problem in P2P lending market, we compare the results of IEFSVM with those of seven state-of-the-art algorithms including the cost-sensitive adaptive boosting (cs-AdaBoost) [46], cost-sensitive Random Forest (cs-RF) [47], EasyEnsemble [48], random under-sampling boosting (RUSBoost) [49], weighted extreme learning machine (w-ELM) [50], cost-sensitive extreme gradient boosting (cs-XGBoost) [7], and EFSVM [12] based on three measures: the area under the receiver operating characteristic curve (AUC), precision, and predicted negative condition rate. AUC has been utilized as an general evaluation criterion of imbalanced classification [51], whereas precision and predicted negative condition rate are additional ratios to check whether an imbalanced classification exhibits a decent performance in investment decision. Then, in order to apply the result of IEFSVM for the investment decision, we adopt the concept of top decile [52] by focusing on the top 10% of instances predicted as most likely to yield high investment returns. Originally, this measure is often used to compare the performance of searching minority instances in churn prediction [53], but we modify the top decile to measure

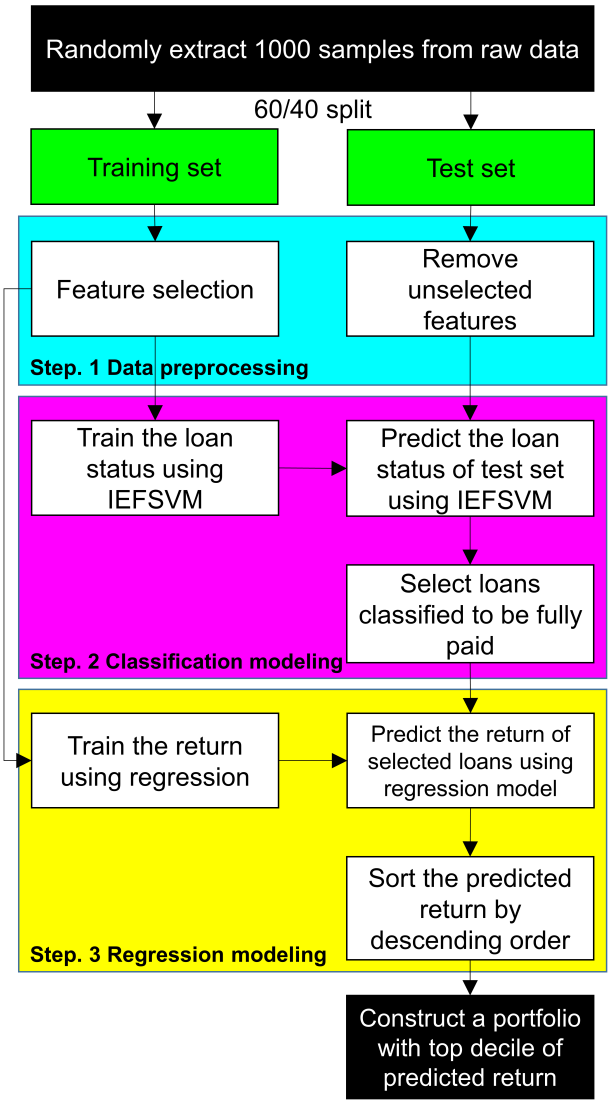

FIGURE 1. Flowchart of proposed investment decision model.

whether it is a good investment decision. The models of Serrano-Cinca and Gutiérrez-Nieto [27] and Guo *et al.* [28] are compared with the proposed model based on investment return and the Sharpe ratio [29] for the loans with the top 10% of the predicted return. Note that the Sharpe ratio is the average return earned per unit of total risk. Therefore, the higher the value of the Sharpe ratio, the more attractive the risk-adjusted return in the investment model. In this context, the performances of our models are compared with both existing classifiers and profitable investment decision models.

III. METHOD

In this section, we first explain the existing EFSVM and compare with the suggested IEFSVM in parallel. Then, we propose an investment decision model with IEFSVM to P2P lending market. The step-by-step scenario of proposed investment decision model is summarized in Fig.1.

A. ENTROPY FUZZY SUPPORT VECTOR MACHINE (EFSVM)

As previously mentioned, we employ IEFSVM in [26], an advanced model of EFSVM, for loan prediction in P2P lending market. Many literatures have proposed methods to compute the fuzzy membership of FSVM; Fan *et al.* [12] suggested to utilize the EFSVM by incorporating the nearest neighbors entropy for imbalanced dataset.

1) ENTROPY FUZZY MEMBERSHIP

Given the training samples $\{x_i, y_i\}_{i=1}^{N}$ with n-dimensional sample x_i, suppose there is a binary classification problem with $y_i \in \{-1, +1\}$. In this case, since we only consider imbalanced dataset, $y_i = +1$ shows that the sample x_i belongs to the minority class, else belongs to the majority class. The EFSVM assigns the fuzzy membership by using the entropy value of each sample based on the FSVM. Entropy of a sample x_i, E_i, is a measure of information's certainty, and applying the entropy to a binary classification yields the following equation.

$$E_i = \begin{cases} 0 & \text{if } p_i = 0 \text{ or } q_i = 0 \\ -p_i \ln(p_i) - q_i \ln(q_i) & \text{otherwise} \end{cases} \quad (1)$$

where p_i and q_i are the probabilities that sample x_i belongs to positive and negative class, respectively. If p_i and q_i are not significantly different, the sample does not know which class it belongs to. Thus, the information is uncertain with a high entropy. On the other hand, if two probabilities are different, it is easy to realize which class the sample belongs to. Hence, the information is reliable with a low entropy.

There are many techniques to obtain two probabilities. In this study, we employ the concept of nearest neighbor. At first, we search k nearest neighbors for each sample. Then, we examine which class the nearest neighbors belong to. Let N_i^+ and N_i^- denote the number of nearest neighbors belonging to the positive and negative class, respectively, then the positive and negative probabilities are $p_i = N_i^+/k$ and $q_i = N_i^-/k$, respectively. Thus, the entropy for each sample can be calculated from the above equation, called nearest neighbors entropy. The high entropy indicates that the information is not clear; the corresponding fuzzy membership should be low. On the contrary, the low entropy indicates that the information is certain; the corresponding fuzzy membership should be high. Therefore, there is a negative relationship between nearest neighbors entropy and fuzzy membership. Considering this relationship, following equation, s_i, is an example of entropy fuzzy membership of a sample x_i [12].

$$s_i = \begin{cases} 1 & \text{if } y_i = +1 \\ (1 - E_i)/\rho & \text{if } y_i = -1. \end{cases} \quad (2)$$

where $\rho = N_{maj}/N_{min}$. Note that N_{maj} and N_{min} are the number of samples in the majority and minority class, respectively. Imbalanced ratio, ρ, is a measure of how imbalanced the number of samples is. It is applied to fuzzy membership to set a low importance to the majority sample. The term $(1 - E_i)$ demonstrates the negative relation between nearest

neighbors entropy and fuzzy membership. The main procedure of EFSVM's fuzzy membership evaluation is outlined in Algorithm 1.

Algorithm 1 Fuzzy Membership Evaluation of EFSVM

Input : Training data $X = \{(x_i, y_i)\}_{i=1}^{N}, y_i \in \{+1, -1\}$, kernel function, neighborhood size k
Output : Fuzzy membership of each instance $\{s_i\}_{i=1}^{N}$

```
Tune neighborhood size
for k in (1, 3, 5, 7, 9, 11, 13, 15) do
    P ← 5-fold of training data {(x_i, y_i)}_{i=1}^N
    for b = 1 to 5 do
        V ← P_{j=b} (Validation set of 5-fold)
        T ← P_{j≠b} (Training set of 5-fold)
        for i = 1 to length(T) do
            if y_i = −1 then
                Find k nearest neighbors for each sample i in T
                Calculate E_i for each sample i in T by Eq.(1)
            end
            Calculate s_i for each sample i in T by Eq.(2)
        end
        Mdl←Fit EFSVM with s_i
        Predict y_i in V with Mdl
    end
    Err_k ← 5-fold CV error of EFSVM with s_i
end
k_opt ← argmin_k Err_k
Evaluate Fuzzy membership
for i = 1 to N do
    if y_i = −1 then
        Find k_opt nearest neighbors for each sample i in X
        Calculate E_i for each sample i in X by Eq.(1)
    end
    Calculate s_i for each sample i in X by Eq.(2)
end
return {s_i}_{i=1}^N
```

According to Algorithm 1, EFSVM chooses a value of k that guarantees the highest accuracy for all data, so that entropy fuzzy membership varies with the number of nearest neighbors, k. A small k has a risk of overfitting since the entropies are obtained only with the near sample, whereas a large k can incur the ignorance of small amounts of information and difficulties of considering a complex distribution since the entropies are calculated rather away from the sample. Therefore, it is important to determine the appropriate number of nearest neighbors for each dataset.

2) QUADRATIC OPTIMIZATION PROBLEM OF EFSVM

In Algorithm 1, the entropy fuzzy membership is decided based on s_i in Eq.(2). The quadratic optimization problem with training set $T = \{(x_i, y_i, s_i) : i = 1, \ldots, N\}$ for FSVM is as follows [54].

$$\begin{aligned} &\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} s_i\varepsilon_i \\ &s.t.\ y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \varepsilon_i, \\ &\quad \varepsilon_i \geq 0, \quad 0 \leq s_i \leq 1, \quad with \quad i = 1, 2, \ldots, N \end{aligned} \tag{3}$$

where w, C, ε_i, $\phi(x)$, and b are the weight for the hyperplane in the feature space, the regularization parameter, the soft error denoting non-negative slack variable of x_i, the non-linear feature mapping, and the bias, respectively. Note that C, only free parameter in the formulation of SVM, is tuned by the parameter selection to balance between classification violation and margin maximization [54].

FSVM determines the hyperplane by multiplying membership (s_i) to soft error (ε_i) to adjust the weight. Also, it has the same result with SVM if the membership has a value of 1 for all i [55]. In other words, the difference between FSVM and SVM depends on the fuzzy membership. Based on the Lagrange multiplier to solve the optimization problem, the equation is transformed into the following dual problem [54].

$$\begin{aligned} &\max \ \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \\ &s.t.\ \sum_{i=1}^{N} y_i\alpha_i = 0, \quad 0 \leq \alpha_i \leq s_i C, \\ &\quad with \quad i = 1, 2, \ldots, N \end{aligned} \tag{4}$$

where $K(x_i, x_j) = \phi(x_i)^T\phi(x_i)$, and α_i denotes the Lagrange multiplier. Then, Sequential Minimal Optimization (SMO) [56] is applied to solve Eq.(4), which yields the optimal value for α_i. Thus, the weight vector is $w = \sum_{i=1}^{N}\alpha_i y_i\phi(x_i)$, whereas the decision function is $f(x) = sign(\sum_{i=1}^{N}\alpha_i y_i K(x, x_i) + b)$ [57].

B. INSTANCE-BASED ENTROPY FUZZY SUPPORT VECTOR MACHINE (IEFSVM)

EFSVM tunes the number of nearest neighbors, k, to maximize the prediction accuracy. However, the disadvantage of this method lies on a unified k for all instances. In this paper, we suggest to utilize IEFSVM proposed in [26] to reflect the change of entropy, which varies in response to different k, in determination of fuzzy membership. The determination of such fuzzy membership is based on the following diversity pattern of nearest neighbors entropy.

1) DIVERSITY PATTERN OF NEAREST NEIGHBORS ENTROPY

To examine the diversity pattern of the nearest neighbors entropy, we provide two examples of the process of calculating the entropy. The first example is for a fixed data point with varying neighborhood size. Fig.2 searches 15 nearest neighbors for a data point i where the neighbors belonging to

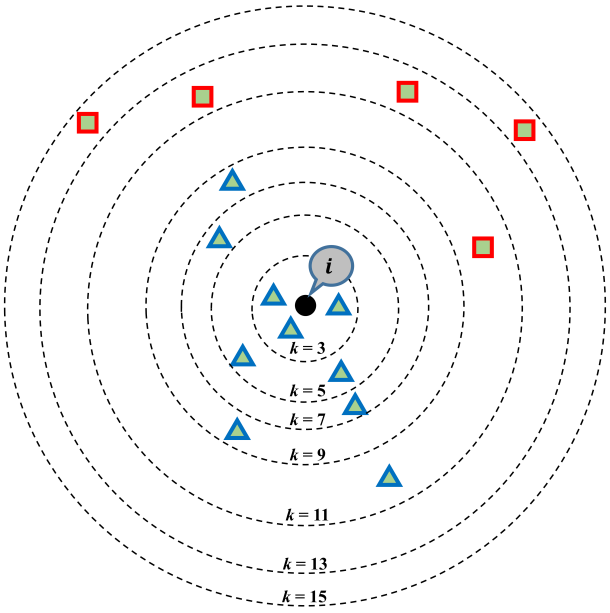

FIGURE 2. Description of nearest neighbors entropy for a fixed data point with varying neighborhood size.

TABLE 1. Description of nearest neighbors entropy for a fixed neighborhood size.

N_i^+	p_i	E_i^9
0	0	0
1	$\frac{1}{9}$	$-\frac{1}{9}\ln(\frac{1}{9})-\frac{8}{9}\ln(\frac{8}{9})=0.3488$
2	$\frac{2}{9}$	$-\frac{2}{9}\ln(\frac{2}{9})-\frac{7}{9}\ln(\frac{7}{9})=0.5297$
3	$\frac{3}{9}$	$-\frac{3}{9}\ln(\frac{3}{9})-\frac{6}{9}\ln(\frac{6}{9})=0.6365$
4	$\frac{4}{9}$	$-\frac{4}{9}\ln(\frac{4}{9})-\frac{5}{9}\ln(\frac{5}{9})=0.6870$
5	$\frac{5}{9}$	$-\frac{5}{9}\ln(\frac{5}{9})-\frac{4}{9}\ln(\frac{4}{9})=0.6870$
6	$\frac{6}{9}$	$-\frac{6}{9}\ln(\frac{6}{9})-\frac{3}{9}\ln(\frac{3}{9})=0.6365$
7	$\frac{7}{9}$	$-\frac{7}{9}\ln(\frac{7}{9})-\frac{2}{9}\ln(\frac{2}{9})=0.5297$
8	$\frac{8}{9}$	$-\frac{8}{9}\ln(\frac{8}{9})-\frac{1}{9}\ln(\frac{1}{9})=0.3488$
9	1	0

the positive and negative classes are represented by triangles and squares, respectively.

According to Fig. 2, 10 out of 15 nearest neighbors are assigned to positive class. Based on Eq.(1), the nearest neighbors entropy is $E_i = -\frac{10}{15}\ln(\frac{10}{15}) - \frac{5}{15}\ln(\frac{5}{15}) = 0.6365$. Likewise, the nearest neighbors entropy can be computed by altering the size of neighborhood. For instance, the nearest neighbors entropies of the neighborhood sizes of 13 and 11 are $E_i = -\frac{10}{13}\ln(\frac{10}{13}) - \frac{3}{13}\ln(\frac{3}{13}) = 0.5402$ with three positive elements and $E_i = -\frac{10}{11}\ln(\frac{10}{11}) - \frac{1}{11}\ln(\frac{1}{11}) = 0.3046$ with one positive element, respectively. Note that the neighborhood sizes of 1, 3, 5, 7, 9 yield zero positive element where $E_i = 0$.

The second example in Table 1 is for a fixed neighborhood size. Let the neighborhood size be 9. Then, the number of elements belonging to the positive class, N_i^+, can range from 0 to 9. Accordingly, the probabilities that sample x_i belongs to positive class, p_i, are $\{0, \frac{1}{9}, \frac{2}{9}, \cdots, \frac{8}{9}, 1\}$, and the corresponding nearest neighbors entropies of sample x_i are specified in the third column of Table 1. Therefore, according to the above two examples, it is shown that the nearest neighbors entropy depends on data point and neighborhood size.

Let k be $(1, 3, 5, 7, 9, 11, 13, 15)$ and E_i^k be nearest neighbors entropy of data point i with k nearest neighbors, then the corresponding entropy pairs, $\{E_i^1, E_i^3, E_i^5, E_i^7, E_i^9, E_i^{11}, E_i^{13}, E_i^{15}\}$, can be calculated. The entropy pairs vary with respect to the class ratio of each neighborhood size. Table 2 depicts all the entropy pairs generated as the neighborhood size changes. In Table 2 from the second to the ninth column, the entropies for each neighborhood size are calculated where the number of nearest neighbors belonging to the positive and negative class, (N_i^+, N_i^-), is written in parentheses.

For the sake of clarity, we also specify the case of Fig.2, located on the second row from the bottom. If we count all different cases of entropy pairs with the set of neighborhood size $\{1, 3, 5, 7, 9, 11, 13, 15\}$, there are total 4374 cases for the following reasons. When k is one, the neighbor can belong to positive class or negative class, which yields the total number of 2 cases. If the neighborhood size increases to 3, there are three additional cases, which are 0, 1, 2 additional positive elements. For instance, (1 positive neighbor, 0 negative neighbor) when k is one can be transformed into the following three cases when k increases to 3: (3 positive neighbors, 0 negative neighbor), (2 positive neighbors, 1 negative neighbor), (1 positive neighbor, 2 negative neighbors). This example is expressed in Fig.3. Such phenomenon indicates that there are $2 \times 3 = 6$ cases of entropy pairs with the set of neighborhood size $\{1, 3\}$. In this respect, there are $2 \times 3^7 = 4374$ cases of entropy pairs with the set of neighborhood size $\{1, 3, 5, 7, 9, 11, 13, 15\}$. As the neighborhood size increases for each data point, the number of nearest neighbors in a class monotonically increases. Specifically, the average and standard deviation of entropies for each data point i are defined as in Eq.(5). Note that the tenth and eleventh columns in Table 2 demonstrate the average and the standard deviation, respectively.

$$\mu_i = \sum_{k=1}^{8} E_i^{2k-1}/8, \quad \sigma_i = \Big(\sum_{k=1}^{8}(E_i^{2k-1} - \mu_i)^2/7\Big)^{\frac{1}{2}} \tag{5}$$

Then, for all 4374 samples, we can construct a scatterplot (μ_i, σ_i) as illustrated in Fig.4a. The horizontal and the vertical axes of the scatterplot are the average and standard deviation

TABLE 2. Nearest neighbors entropy according to neighborhood size (Note that: N_i^+ and N_i^- are written in parentheses).

i	E_i^1	E_i^3	E_i^5	E_i^7	E_i^9	E_i^{11}	E_i^{13}	E_i^{15}	μ_i	σ_i
1	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0 (0, 13)	0 (0, 15)	0	0
2	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0 (0, 13)	0.2449 (1, 14)	0.0306	0.0866
3	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0 (0, 13)	0.3927 (2, 13)	0.0491	0.1388
4	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.2712 (1, 12)	0.2449 (1, 14)	0.0645	0.1197
5	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.2712 (1, 12)	0.3927 (2, 13)	0.0830	0.1570
6	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.2712 (1, 12)	0.5004 (3, 12)	0.0964	0.1888
7	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.4293 (2, 11)	0.3927 (2, 13)	0.1027	0.1905
8	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.4293 (2, 11)	0.5004 (3, 12)	0.1162	0.2160
9	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.4293 (2, 11)	0.5799 (4, 11)	0.1262	0.2370
10	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0.3046 (1, 10)	0.2712 (1, 12)	0.2449 (1, 14)	0.1026	0.1425
11	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0.3046 (1, 10)	0.2712 (1, 12)	0.3927 (2, 13)	0.1211	0.1704
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Fig.2.	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0.3046 (10, 1)	0.5402 (10, 3)	0.6365 (10, 5)	0.1852	0.2714
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4374	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0 (13, 0)	0 (15, 0)	0	0

of entropies, respectively. Since this scatterplot seems to fit well with polar coordinates, we transform (μ_i, σ_i) into polar coordinates as in Eq.(6).

$$d_i = (\mu_i^2 + \sigma_i^2)^{\frac{1}{2}}, \quad \theta_i = \tan^{-1}(\mu_i/\sigma_i) \tag{6}$$

In Fig.4a, groups of comparable θ_i can be discovered by moving the red line (i.e. increasing θ_i). Then, the scatterplot can be seen as a combination of line segments leading to the origin. Therefore, it is necessary to analyse the common between lines and difference BETWEEN line segments. Then, we examine the number of nonzero entropies of each entropy pair. In Fig.2, the entropy pair is (0, 0, 0, 0, 0, 0.3046, 0.5402, 0.6365), which yields three nonzero entropies (i.e. 0.3046, 0.5402, and 0.6365). In this circumstance, data points with one and two nonzero entropies can be considered as specified in Table 3.

Table 3 only depicts the entropy pairs that possess one or two nonzero entropies among the data points of Table 2. Note that $i = 2, 3, 4372, 4373$ are the entropy pairs of one nonzero entropy, whereas $i = 4, 5, \ldots, 9, 4366, 4367, \ldots, 4371$ are those of two nonzero entropies. For one nonzero entropies, all nearest neighbors belong to the positive class up to 13, whereas the neighborhoods of negative class appear when the neighborhood size becomes 15. It indicates that $E_i^k = 0$ for $k = 1, 3, 5, 7, 9, 11, 13$, and $E_i^{15} \neq 0$. For two nonzero entropies, all nearest neighbors belong to the positive class up to 11, whereas the neighborhoods of negative class appear when the neighborhood size is over 11. It refers that $E_i^k = 0$ for $k = 1, 3, 5, 7, 9, 11$, and $E_i^k \neq 0$ for $k = 13, 15$. Then, we classify (d_i, θ_i) by the number of nonzero entropies, which is shown in Fig.4b. Fig.4b categorizes the data points of Fig.4a scatterplot from one to six nonzero entropies. Then, the data points having the same number of nonzero entropies can be seen as a line. In other words, the data points of the polar coordinates can be classified according to the number of nonzero entropies. We also transform Fig.4b into polar coordinates as in Fig.4c. Then, as expected, the data points having the same number of nonzero entropies have the same θ_i.

To further examine the scatterplot, we construct the following two cases. The first is the fixed θ_i and increase of d_i. In this case, as in Fig.4c, the number of nonzero entropy is fixed. Then, when d_i increases, μ_i and σ_i proportionally increase by Fig.4b. The increase in μ_i refers to the increase of entropy, which also indicates that the information is uncertain. Furthermore, the increment of σ_i indicates that the components of entropy pairs highly vary according to the neighborhood size, and likewise the information is uncertain.

The second condition is an increment of θ_i. When θ_i becomes larger, d_i proportionally increases as in Fig.4c. As mentioned above, increase in d_i indicates uncertain information. Also, increase in θ_i incurs an increment of the number of nonzero entropies. The increment of the number of nonzero entropies indicates an increment of both μ_i and σ_i in general. Therefore, it refers to the uncertain information. In overall, the increases in d_i and θ_i yield the uncertainty of

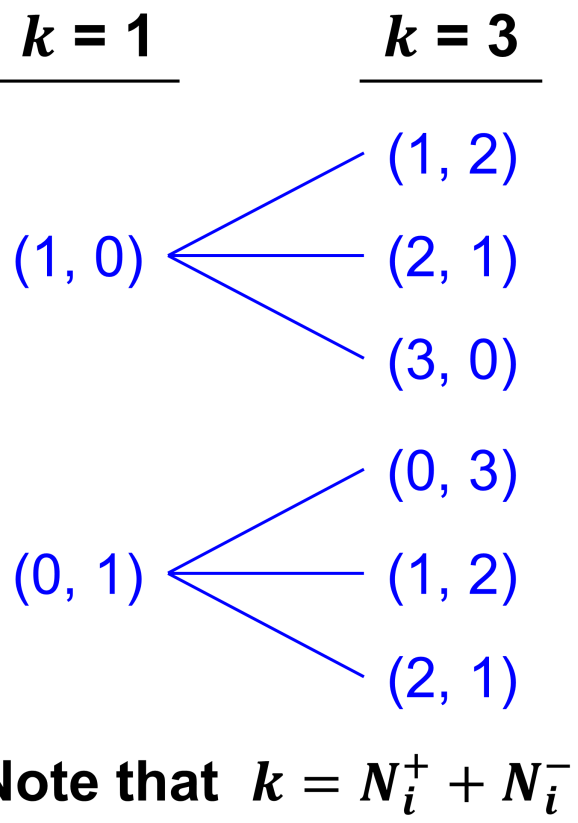

FIGURE 3. Description of (N_i^+, N_i^-) in the case for the set of neighborhood size {1, 3}.

information. Through the graphical analysis of nearest neighbors entropy, we found that more plausible usages of entropy can be obtained when new fuzzy membership can be well defined.

2) INSTANCE-BASED ENTROPY FUZZY MEMBERSHIP

Based on the diversity pattern of nearest neighbors entropy, the four logics for decision of s_i are considered as follows. At first, as in Eq.(2), s_i should be reduced if μ_i increases since the entropy and fuzzy membership exhibit a negative relationship. Secondly, if σ_i increases, s_i also should be decreased. A high σ_i implies a large variance between entropies in each entropy pair. Also, the number of neighbors belonging to a class varies by neighborhood size, which yields the uncertain information. Thirdly, when θ_i increases, s_i should be reduced. Lastly, when d_i increases, s_i should decrease. Note that the high d_i and θ_i lead to high μ_i and σ_i according to the graphical analysis in Fig.4, which also causes the failure of providing reliable information for imbalanced classification.

The above four logics suggest that s_i should decrease when d_i and θ_i increase. Based on this idea, the instance-based entropy fuzzy membership incorporating the nearest neighbors entropy information can be defined as follows [26].

$$s_i = \begin{cases} 1 & \text{if } y_i = +1 \\ (1 - \dfrac{d_i\theta_i - \min_i d_i\theta_i}{\max_i d_i\theta_i - \min_i d_i\theta_i})/\rho & \text{if } y_i = -1. \end{cases} \quad (7)$$

where $\min_i d_i\theta_i$ and $\max_i d_i\theta_i$ denote the minimum and maximum values of $d_i\theta_i$, respectively. When constructing a training set, a certain portion of the instances is used. For example, 600 instances of 60% are selected as training sets after randomly selecting 1000 instances. At this time, if the training set is scatter-plotted when d_i and θ_i are obtained with 600 training sets by Eq.(6), it will be a subset of the points in Fig 4a. In other words, the points obtained for d_i and θ_i with the training set may not represent all the points in Fig.4a. Therefore, for the 600 training samples, the maximum and minimum values of $d_i \times \theta_i$ will vary with each training set. IEFSVM determines the fuzzy membership by Eq.(7) where the main process of the fuzzy membership evaluation is in Algorithm 2. Also, we consider the level curve of the fuzzy membership to visualize IEFSVM. Fig.4d whose curve is $d_i \times \theta_i = l$ for $l = 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75$ is supplement of the scatter plot in Fig.4a.

Algorithm 2 Fuzzy Membership Evaluation of IEFSVM

Input : Training data $X = \{(x_i, y_i)\}_{i=1}^N$, $y_i \in \{+1, -1\}$, kernel function, neighborhood size k
Output : Fuzzy membership of each instance $\{s_i\}_{i=1}^N$

```
Evaluate Fuzzy membership
for i = 1 to N do
    if y_i = −1 then
        for k in (1, 3, 5, 7, 9, 11, 13, 15) do
            Find k nearest neighbors for each sample i in X
            Calculate E_i for each sample i in X by Eq.(1)
        end
        Calculate μ_i and σ_i for each sample i in X by Eq.(5)
        Calculate d_i and θ_i for each sample i in X by Eq.(6)
    end
    Calculate s_i for each sample i in X by Eq.(7)
end
return {s_i}_{i=1}^N
```

Both the EFSVM and IEFSVM are divided into two steps to calculate the complexity of the algorithms, which are searching nearest neighbors of each sample and learning the classifiers. Since the process of finding nearest neighbors for each sample is a common process for both algorithms, there is no difference in complexity as $O(kNN_{neg})$ [12], where k, N, and N_{neg} denote the number of nearest neighbors, the number of training samples, and the number of negative samples. For minority samples, fuzzy membership is set to be 1. Therefore, the searching for nearest neighbors is not required. IEFSVM can calculate the support vector only once; nevertheless, EFSVM should learn the model for all neighborhood sizes to tune the number of nearest neighbors. Therefore, the time to evaluate the support vector of IEFSVM is shorter than that of EFSVM by the number of neighborhood sizes.

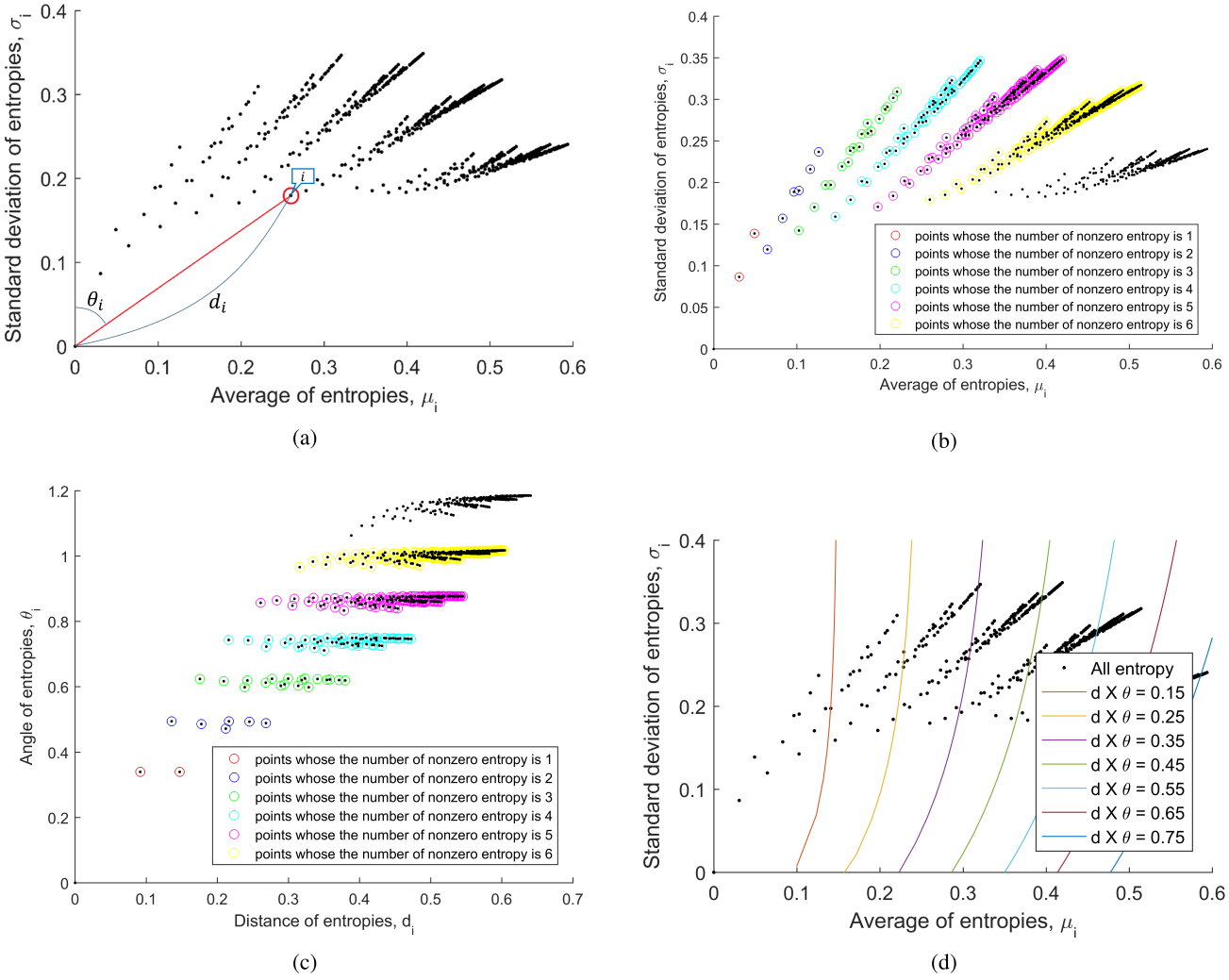

FIGURE 4. (a) Scatterplot of average and standard deviation of all nearest neighbors entropy (b) Scatterplot by the number of nonzero nearest neighbors entropy (c) Polar coordinate of the scatterplot by the number of nonzero nearest neighbors entropy (d) Scatterplot and level curve of the fuzzy membership.

C. INVESTMENT DECISION MODEL

Utilization of SVM-based classifier to imbalanced dataset assigns a high weight to the minority class, which eventually yields a biased minority data against the decision surface. On the other hand, the majority data spread regardless of the decision surface due to its low importance. In this case, we obtain very high percentage of majority data in selecting the samples predicted by the majority class. Applying such idea to P2P data, we can select fully paid loans at a very high rate. Therefore, the first step is to select the loans that are expected to be fully paid based on IEFSVM.

Serrano-Cinca and Gutiérrez-Nieto [27] showed that a portfolio based on a simple regression can achieve a high investment return with the loans that are expected to yield a high return. Likewise, in the study, we propose a multiple linear regression on returns where explanatory variables are the same as those used for IEFSVM. Then, we predict and rank the returns of loans in the test set through the predicted return. A simple two-step mechanism for the proposed investment model is as follows. At first, the model includes the loans that are predicted to be fully paid based on IEFSVM as a pool. Then, the model selects the top 10% of the loans from the pool that are expected to obtain a high investment return based on the multiple regression. Note that the final portfolio is invested with the equal amount of money for all selected loans.

IV. EMPIRICAL RESULTS

In this section, we first demonstrate the imbalanced characteristics of P2P lending data. Then, we apply several imbalanced classification models including IEFSVM and select the loans predicted to be fully paid. After regressing on returns for the loans, we compose a portfolio to yield a high expected returns based on the proposed investment model.

A. DATASET DESCRIPTION

The data are obtained from the Lending Club. The Lending Club is the largest intermediary for connecting P2P

TABLE 3. Nearest neighbors entropy with one and two nonzero values (Note that: N_i^+ and N_i^- are written in parentheses).

i	E_i^1	E_i^3	E_i^5	E_i^7	E_i^9	E_i^{11}	E_i^{13}	E_i^{15}	μ_i	σ_i
2	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0 (0, 13)	0.2449 (1, 14)	0.0306	0.0866
3	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0 (0, 13)	0.3927 (2, 13)	0.0491	0.1388
4372	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0 (13, 0)	0.3927 (13, 2)	0.0491	0.1388
4373	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0 (13, 0)	0.2449 (14, 1)	0.0306	0.0866
4	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.2712 (1, 12)	0.2449 (1, 14)	0.0645	0.1197
5	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.2712 (1, 12)	0.3927 (2, 13)	0.0830	0.1570
6	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.2712 (1, 12)	0.5004 (3, 12)	0.0964	0.1888
7	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.4293 (2, 11)	0.3927 (2, 13)	0.1027	0.1905
8	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.4293 (2, 11)	0.5004 (3, 12)	0.1162	0.2160
9	0 (0, 1)	0 (0, 3)	0 (0, 5)	0 (0, 7)	0 (0, 9)	0 (0, 11)	0.4293 (2, 11)	0.5799 (4, 11)	0.1262	0.2370
4366	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0.4293 (11, 2)	0.5799 (11, 4)	0.1262	0.2370
4367	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0.4293 (11, 2)	0.5004 (12, 3)	0.1162	0.2160
4368	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0.4293 (11, 2)	0.3927 (13, 2)	0.1027	0.1905
4369	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0.2712 (12, 1)	0.5004 (12, 3)	0.0964	0.1888
4370	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0.2712 (12, 1)	0.3927 (13, 2)	0.0830	0.1570
4371	0 (1, 0)	0 (3, 0)	0 (5, 0)	0 (7, 0)	0 (9, 0)	0 (11, 0)	0.2712 (12, 1)	0.2449 (14, 1)	0.0645	0.1197

loans, and it determines and provides the rating of the loans based on the borrower's personal information and financial data such as annual income, employment length, interest rate, number of open account, revolving utilization rate, and et cetera.

1) LENDING CLUB GRADE

The Lending Club classifies each borrower's financial grade based on the borrower's financial information. The grades consist of 7 elements from A to G. More stable the borrower is, more likely he/she receives grade A. Thus, the interest rate increases from A to G grades. According to the Lending Club, the loan statistics by ratings are shown in Table 4.

TABLE 4. Loan statistics by LC grade.

Grade	A	B	C	D	E	F+G	All
Loan amount($\times 10^8$\$)	9.56	14.60	9.86	4.90	1.34	0.33	40.58
Interest rate(%)	7.52	11.52	14.57	17.61	20.46	23.75	12.64
Historical return(%)	7.12	9.83	10.05	10.62	10.68	11.59	9.45
Standard deviation(%)	13.25	18.92	24.44	29.15	33.27	37.59	21.92

In Table 4, the total amount of loans used in this study is roughly 4 billion dollars, of which the loan amount of B grade is the largest with 1.460 billion. As the interest rate steadily increases from A to G grades, the historical return also rises consistently. At high grades, the return is low due to low interest rates. At low grades, despite the high interest rate, the return is low due to high default rates. The standard deviations are much higher than returns for all grades, which yields the Sharpe ratio less than one. Note that practically speaking, the portfolio return of loans can be dramatically improved if we can detect loans to be fully paid at lower grades.

2) DISTRIBUTION OF RETURN

The histogram of the creditor's return is shown in Fig.5. Histogram in above includes all loans, whereas histogram in below only shows the negative returns. Obviously, the proportion of the loans with positive returns is much larger than that with negative returns. Out of 331,878 loans, 288,398 loans are fully paid, accounting for 86.9%. In addition, the loans with the positive returns are 292,851, accounting for 88.2%. Therefore, the imbalanced classification is suitable for solving the loan status prediction problem.

3) EXPLANATORY DATA ANALYSIS

Specifically, we use the data from 2007 to 2014 for 3-year loans. Note that only the "Fully paid" and "Charged off" loans are selected, which yields 331,878 samples in total.

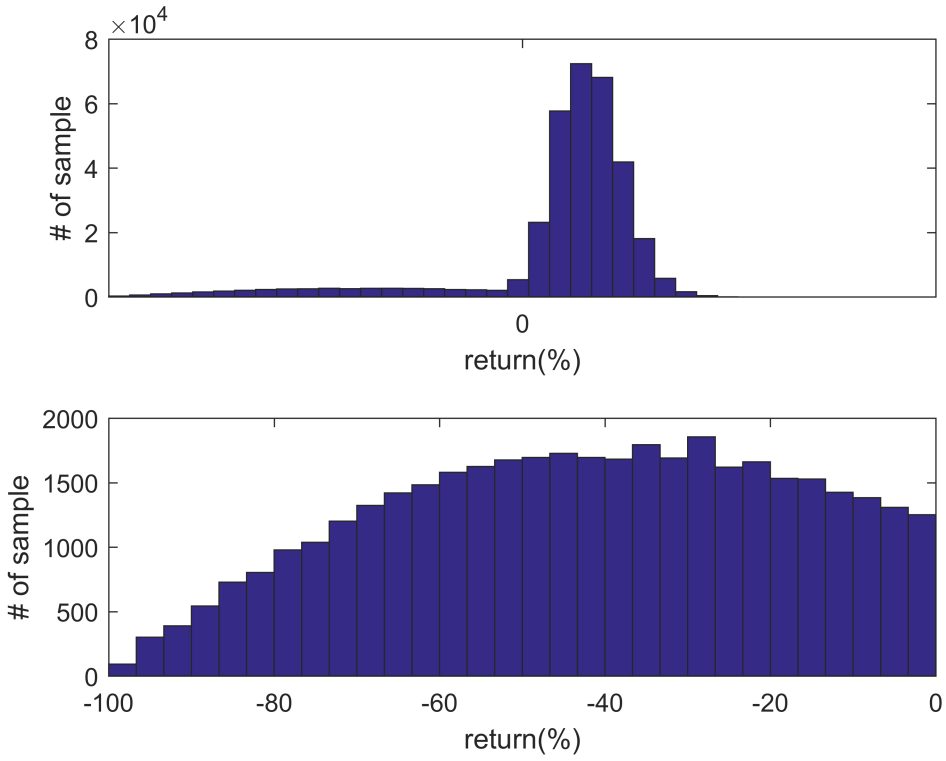

FIGURE 5. Histogram of the creditor's return.

The loan status is used for the imbalanced classification model as the response variable, whereas the return is used for the regression model. Then, we select independent variables from the pool of 128 variables provided by the Lending Club data. Note that numerical variables are normalized such that the maximum and minimum values are 1 and -1, respectively, whereas categorical variables are binarized to obtain dummy variables. Then, applying the gradient boosting method [58] to measure the importance of each variable, we finally select the independent variables with high importance.

B. EXPERIMENTAL SET-UP

1) PARAMETERS FOR CLASSIFICATION METHODS

For comparison, we evaluate the performance of IEFSVM against cs-AdaBoost, cs-RF, EasyEnsemble, RUSBoost, w-ELM, cs-XGBoost, and EFSVM. For SVM-based learning machines such as EFSVM and IEFSVM, the radial basis function (RBF) kernel or linear kernel can be used. The regularization parameter C is chosen from $\{2^{-6}, 2^{-4}, \ldots, 2^{4}, 2^{6}\}$. To calculate the entropy, the number of nearest neighbors is selected from $\{1, 3, 5, 7, 9, 11, 13, 15\}$. For tree-based learning machines such as cs-AdaBoost, cs-RF, EasyEnsemble, and RUSBoost, we select 100 maximum learning iterations, whereas tuning of cs-XGBoost follows the methods in Xia *et al*. [7] and Jain [59]. Procedures for tuning the parameters are performed through a 5-fold cross validation.

2) PERFORMANCE MEASURES

We provide three performance measures for imbalanced classification and two measures for investment decision model. Table 5 is a confusion matrix that visualizes the classification performance.

TABLE 5. Confusion matrix.

Total population		Predicted	
		Fully paid	Charged off
Actual	Fully paid	True Fully paid	False Charged off
	Charged off	False Fully paid	True Charged off

Then, AUC [60] is used for comparison measure of the classification performance of each learning machine. AUC can be defined as follows.

$$AUC = (1 + TP_{rate} - FP_{rate})/2. \tag{8}$$

where TP_{rate} and FP_{rate} denote the ratio of positive samples correctly classified and that of negative samples misclassified, respectively.

Since the investment decision proposed in this paper discards loans that are predicted to be default, we define precision as follows.

$$\text{Precision} = \text{True Fully paid} \,/\, \text{Predicted Fully paid} \tag{9}$$

TABLE 6. Performance measures for all classifiers.

	cs-AdaBoost	cs-RF	EasyEnsemble	RUSBoost	w-ELM	cs-XGBoost	EFSVM	IEFSVM
AUC	59.23 ± 3.95	52.58 ± 2.63	55.28 ± 3.95	58.37 ± 5.07	56.52 ± 3.64	56.62 ± 4.15	57.14 ± 3.85	59.38 ± 2.85
	2	8	7	3	6	5	4	1
Precision	91.98 ± 2.59	87.63 ± 0.65	89.20 ± 1.65	91.39 ± 2.81	89.47 ± 1.42	89.26 ± 1.48	90.42 ± 2.13	92.16 ± 1.83
	2	8	7	3	5	6	4	1
Predicted negative condition rate	52.04 ± 19.26	7.49 ± 1.84	45.37 ± 5.36	50.24 ± 20.12	39.45 ± 6.18	31.73 ± 6.93	47.61 ± 13.85	57.28 ± 8.85
	2	8	5	3	6	7	4	1
Return with top 10%	10.06 ± 4.63	12.26 ± 5.51	11.91 ± 4.93	10.34 ± 4.48	11.44 ± 5.34	11.58 ± 5.00	12.63 ± 5.59	14.99 ± 2.87
	8	3	4	7	6	5	2	1
Return / Standard dev.	2.175	2.226	2.417	2.31	2.145	2.319	2.26	5.227
	7	6	2	4	8	3	5	1

TABLE 7. Significance tests in classifiers.

Pairwise comparison	AUC		Precision		Predicted negative condition rate		Return with top 10%	
	p	Hypothesis (0.05)	p	Hypothesis (0.05)	p	Hypothesis (0.05)	p	Hypothesis (0.05)
IEFSVM vs. cs-AdaBoost	0.1651	Not rejected	0.0364	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**
IEFSVM vs. cs-RF	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**
IEFSVM vs. EasyEnsemble	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**
IEFSVM vs. RUSBoost	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**
IEFSVM vs. w-ELM	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**
IEFSVM vs. cs-XGBoost	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**
IEFSVM vs. EFSVM	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**	< 0.001	**Rejected**

However, the disadvantage of such investment decision is that the more data are discarded if the higher performance of the classifier is objected. Therefore, we also define predicted negative condition rate to measure the discarded data.

$$\text{Predicted negative condition rate} = \text{Predicted Charged off} \;/\; \text{Total Population} \quad (10)$$

Lastly, we compare the average return and the Sharpe ratio for selected loans in each portfolio to evaluate the investment decisions.

C. ANALYSIS OF EXPERIMENTAL RESULTS

This section demonstrates the performance of IEFSVM and proposed investment decision model. The comparison is based on two types of benchmarks, namely, classifiers adapted to imbalanced dataset and investment models that offer profitable portfolios.

1) COMPARISON OF CLASSIFIERS

For the classfiers, we conduct experiments to compare IEFSVM with seven other classifiers in order to demonstrate the effectiveness of fuzzy membership in our model on imbalanced data. The results are evaluated based on the AUC value, precision, and predicted negative condition rate. Then, we compute the investment return and the Sharpe ratio with top decile of predicted return. Table 6 shows the results of classifiers and specifies the ranks of the measures under each value. Also, we perform a significance test to confirm whether the proposed IEFSVM statistically outperforms the benchmarks as in Table 7.

The results shown in Table 6 reveal that IEFSVM far outperforms other algorithms. For other performance measures except AUC, IEFSVM achieves the best performance. However, as illustrated in Table 7, it does not statistically overwhelm cs-AdaBoost in terms of AUC. Also, IEFSVM significantly outperforms EFSVM for all measures, which supports our approach to enhance EFSVM with the instance-based model. In the meantime, IEFSVM filters the loans the most that are expected to be default. In terms of investment measures, IEFSVM statistically outperforms other algorithms as for the return and the Sharpe ratio in top decile of predicted returns. In detail, cs-AdaBoost shows reasonably high performance on measures of AUC, precision, and predicted negative condition rate, whereas it indicates a poor performance in terms of the return and the Sharpe ratio of top decile. EasyEnsemble, on the contrary, shows a good performance in terms of the return and the Sharpe ratio of top decile, but poor performance for AUC, precision, and predicted negative condition rate. In this respect, IEFSVM can be regarded as a reasonable and stable classifier with the highest performance on all measures.

2) COMPARISON OF THE INVESTMENT MODEL

For the investment models, we compare proposed investment decision model against other models in recent studies. Serrano-Cinca and Gutiérrez-Nieto [27] proposed a profit scoring decision support system by utilizing the internal rate of return (IRR). IRR is the effective interest rate that the lender receives. Hence, it can be used to predict the profitability in P2P lending. Note that we state this method as Benchmark 1. Meanwhile, Guo *et al*. [28] developed an instance-based credit risk assessment model by quantifying the return and risk of each loan. This study also formulated the investment decision model as a portfolio optimization problem with boundary constraints. Note that we state this method as Benchmark 2. Table 8 demonstrates the results of investment returns, standard deviations, and the Sharpe ratio of each model. Also, we perform a significance test to explore whether the proposed investment decision model statistically outperforms the benchmarks as in Table 9.

TABLE 8. Performance measures for all investment models.

	Benchmark1	Benchmark2	IEFSVM
Return(%)	12.475	8.707	14.99
Standard dev.(%)	5.687	2.254	2.868
Return / Standard dev.	2.193	3.863	5.227

TABLE 9. Significance tests in investment decision models.

Pairwise comparison	Investment return	
	p	Hypothesis (0.05)
IEFSVM vs. Benchmark1	< 0.001	**Rejected**
IEFSVM vs. Benchmark2	< 0.001	**Rejected**

The results in Table 8 show that IEFSVM outperforms other investment models. Although IEFSVM has a higher standard deviation than Benchmark 2, it statistically overwhelms two models in terms of the higher returns and the Sharpe ratio as illustrated in Table 9. Therefore, the proposed model based on IEFSVM can enhance the performance of investment decision in P2P lending market.

V. CONCLUSION

This paper proposes investment decision model in P2P lending market by constructing a portfolio based on IEFSVM. We first apply IEFSVM in loan evaluation for P2P lending and select loans that are predicted to be fully paid. We modify EFSVM into instance-based model to cope with its drawback, whose determination of fuzzy membership is only based on a unified neighborhood size. Note that IEFSVM considers the combination of nearest neighbors entropy from near to far distance for each data point in the evaluation of fuzzy membership. In this context, we efficiently reflect all information of each data point when assigning fuzzy membership. To show the effectiveness of IEFSVM, we set EFSVM as one of the imbalanced classifier benchmarks with six other classifiers. Then, we rank and select loans predicted to yield high returns using a multiple regression model, which produces an investment portfolio based on selected loans.

Specifically, in experimental studies, the performance of our approach is compared with two types of benchmarks, namely, the seven imbalanced classifiers and two investment decision models. For imbalanced classifiers, the classification results reveal that IEFSVM statistically outperforms other classifiers including cs-AdaBoost, cs-RF, EasyEnsemble, RUSBoost, w-ELM, cs-XGBoost, and EFSVM in terms of AUC, precision, predicted negative condition rate, returns with top 10%, and the Sharpe ratio. Note that IEFSVM is statistically superior to cs-AdaBoost in most of performance measures except the AUC. Considering that loan status prediction problem in P2P lending market separates the minority class from the majority class, we conclude that the utilization of IEFSVM is successful since it improves the classification performances and well detects fully paid loans. For investment decision model, the performance of proposed model statistically outperforms that of existing investment decision models in terms of return and the Sharpe ratio. Therefore, based on empirical results, we conclude that IEFSVM can be utilized as a decent classification model for P2P lending investment decisions.

Despite its contribution, there are limitations on our model that should be addressed in the future studies. In terms of fuzzy membership, it is possible to develop more elaborative instance-based fuzzy membership by extending the neighborhood size based on the polar coordinate distribution of (d_i, θ_i). Also, we are planning to construct more sophisticated investment decision model to improve a simple multivariate regression.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (No. 2018R1C1B5043835).

REFERENCES

[1] P. Saha, I. Bose, and A. Mahanti, "A knowledge based scheme for risk assessment in loan processing by banks," *Decision Support Syst.*, vol. 84, pp. 78–88, Apr. 2016.

[2] L. Puro, J. E. Teich, H. Wallenius, and J. Wallenius, "Borrower decision aid for people-to-people lending," *Decision Support Syst.*, vol. 49, no. 1, pp. 52–60, 2010.

[3] R. Tsaih, Y.-J. Liu, W. Liu, and Y.-L. Lien, "Credit scoring system for small business loans," *Decision Support Syst.*, vol. 38, no. 1, pp. 91–99, Oct. 2004.

[4] R. T. Stewart, "A profit-based scoring system in consumer credit: making acquisition decisions for credit cards," *J. Oper. Res. Soc.*, vol. 62, no. 9, pp. 1719–1725, 2011.

[5] T. Verbraken, C. Bravo, R. Weber, and B. Baesens, "Development and application of consumer credit scoring models using profit-based classification measures," *Eur. J. Oper. Res.*, vol. 238, no. 2, pp. 505–513, 2014.

[6] S. Maldonado, C. Bravo, J. López, and J. Pérez, "Integrated framework for profit-based feature selection and SVM classification in credit scoring," *Decision Support Syst.*, vol. 104, pp. 113–121, Dec. 2017.

[7] Y. Xia, C. Liu, and N. Liu, "Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending," *Electron. Commerce Res. Appl.*, vol. 24, pp. 30–49, Jul./Aug. 2017.
[8] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Syst. Appl.*, vol. 78, pp. 225–241, Jul. 2017.
[9] W. Li, S. Ding, Y. Chen, and S. Yang, "Heterogeneous ensemble for default prediction of peer-to-peer lending in China," *IEEE Access*, vol. 6, pp. 54396–54406, Mar. 2018.
[10] C. Jiang, Z. Wang, R. Wang, and Y. Ding, "Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending," *Ann. Oper. Res.*, vol. 266, nos. 1–2, pp. 511–529, 2018.
[11] Y. Xia, X. Yang, and Y. Zhang, "A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending," *Electron. Commerce Res. Appl.*, vol. 30, pp. 111–124, Jul./Aug. 2018.
[12] Q. Fan, Z. Wang, D. Li, D. Gao, and H. Zha, "Entropy-based fuzzy support vector machine for imbalanced datasets," *Knowl.-Based Syst.*, vol. 115, pp. 87–99, Jan. 2017.
[13] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
[14] Y. Chen, K. Wu, X. Chen, C. Tang, and Q. Zhu, "An entropy-based uncertainty measurement approach in neighborhood systems," *Inf. Sci.*, vol. 279, pp. 239–250, Sep. 2014.
[15] F. Zhu, J. Yang, C. Gao, S. Xu, N. Ye, and T. Yin, "A weighted one-class support vector machine," *Neurocomputing*, vol. 189, pp. 1–10, May 2016.
[16] F. Zhu, J. Yang, N. Ye, C. Gao, G. Li, and T. Yin, "Neighbors' distribution property and sample reduction for support vector machines," *Appl. Soft Comput.*, vol. 16, pp. 201–209, Mar. 2014.
[17] F. Zhu, N. Ye, W. Yu, S. Xu, and G. Li, "Boundary detection and sample reduction for one-class support vector machines," *Neurocomputing*, vol. 123, pp. 166–173, Jan. 2014.
[18] F. Zhu, J. Yang, J. Gao, and C. Xu, "Extended nearest neighbor chain induced instance-weights for SVMs," *Pattern Recognit.*, vol. 60, pp. 863–874, Dec. 2016.
[19] Y. Chen and Y. Hao, "A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction," *Expert Syst. Appl.*, vol. 80, pp. 340–355, Sep. 2017.
[20] Z. Pan, Y. Wang, and W. Ku, "A new k-harmonic nearest neighbor classifier based on the multi-local means," *Expert Syst. Appl.*, vol. 67, pp. 115–125, Jan. 2017.
[21] J. Gou, Y. Zhan, Y. Rao, X. Shen, X. Wang, and W. He, "Improved pseudo nearest neighbor classification," *Knowl.-Based Syst.*, vol. 70, pp. 361–375, Nov. 2014.
[22] Ö. F. Ertuğrul and M. E. Tağluk, "A novel version of k nearest neighbor: Dependent nearest neighbor," *Appl. Soft Comput.*, vol. 55, pp. 480–490, Jun. 2017.
[23] Y. Zhu, Z. Wang, and D. Gao, "Gravitational fixed radius nearest neighbor for imbalanced problem," *Knowl.-Based Syst.*, vol. 90, pp. 224–238, Dec. 2015.
[24] X. Zhang, Y. Li, R. Kotagiri, L. Wu, Z. Tari, and M. Cheriet, "KRNN: *k* rare-class nearest neighbour classification," *Pattern Recognit.*, vol. 62, pp. 33–44, Feb. 2017.
[25] F. Bulut and M. F. Amasyali, "Locally adaptive *k* parameter selection for nearest neighbor classifier: One nearest cluster," *Pattern Anal. Appl.*, vol. 20, no. 2, pp. 415–425, 2017.
[26] P. Cho, M. Lee, and W. Chang. (2018). "Instance-based entropy fuzzy support vector machine for imbalanced data." [Online]. Available: https://arxiv.org/abs/1807.03933
[27] C. Serrano-Cinca and B. Gutiérrez-Nieto, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending," *Decision Support Syst.*, vol. 89, pp. 113–122, Sep. 2016.
[28] Y. Guo *et al.*, "Instance-based credit risk assessment for investment decisions in P2P lending," *Eur. J. Oper. Res.*, vol. 249, no. 2, pp. 417–426, Mar. 2016.
[29] W. F. Sharpe, "The Sharpe ratio," *J. Portfolio Manage.*, vol. 21, no. 1, pp. 49–58, 1994.
[30] C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios, "Determinants of default in P2P lending," *PLoS ONE*, vol. 10, no. 10, 2015, Art. no. e0139427.
[31] X. Zeng, L. Liu, S. Leung, J. Du, X. Wang, and T. Li, "A decision support model for investment on P2P lending platform," *PLoS ONE*, vol. 12, no. 9, 2017, Art. no. e0184242.
[32] S. Lessmann, B. Baesens, L. C. Thomas, and H.-V. Seow, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, Oct. 2015.
[33] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *J. Oper. Res. Soc.*, vol. 54, no. 6, pp. 627–635, 2003.
[34] G. G. Chen and T. Åstebro, "Bound and collapse Bayesian reject inference for credit scoring," *J. Oper. Res. Soc.*, vol. 63, no. 10, pp. 1374–1387, 2012.
[35] K. Kennedy, B. M. Namee, and S. J. Delany, "Using semi-supervised classifiers for credit scoring," *J. Oper. Res. Soc.*, vol. 64, no. 4, pp. 513–529, 2013.
[36] J. Ouenniche, K. Bouslah, J. M. Cabello, and F. Ruiz, "A new classifier based on the reference point method with application in bankruptcy prediction," *J. Oper. Res. Soc.*, vol. 69, no. 10, pp. 1653–1660, Jan. 2018.
[37] C. Liberati and F. Camillo, "Personal values and credit scoring: new insights in the financial prediction," *J. Oper. Res. Soc.*, vol. 69, no. 12, pp. 1994–2005, 2018.
[38] A. A. Aduenko, A. P. Motrenko, and V. V. Strijov, "Object selection in credit scoring using covariance matrix of parameters estimations," *Ann. Oper. Res.*, vol. 260, nos. 1–2, pp. 3–21, 2018.
[39] Z. Affes and R. Hentati-Kaffel, "Forecast bankruptcy using a blend of clustering and MARS model: Case of US banks," *Ann. Oper. Res.*, pp. 1–38, Apr. 2018.
[40] A. Marqués, V. García, and J. S. Sánchez, "A literature review on the application of evolutionary computing to credit scoring," *J. Oper. Res. Soc.*, vol. 64, no. 9, pp. 1384–1399, 2013.
[41] J. Sun, Z. Shang, and H. Li, " Imbalance-oriented SVM methods for financial distress prediction: a comparative study among the new SB-SVM-ensemble method and traditional methods," *J. Oper. Res. Soc.*, vol. 65, no. 12, pp. 1905–1919, 2014.
[42] A. I. Marqués, V. García, and J. S. Sánchez, "On the suitability of resampling techniques for the class imbalance problem in credit scoring," *J. Oper. Res. Soc.*, vol. 64, no. 7, pp. 1060–1070, 2013.
[43] Z. Chen, T. Lin, X. Xia, H. Xu, and S. Ding, "A synthetic neighborhood generation based ensemble learning for the imbalanced data classification," *Appl. Intell.*, vol. 48, no. 8, pp. 2441–2457, 2018.
[44] Z. Chen, T. Lin, R. Chen, Y. Xie, and H. Xu, "Creating diversity in ensembles using synthetic neighborhoods of training samples," *Appl. Intell.*, vol. 47, no. 2, pp. 570–583, 2017.
[45] S. Ando, "Classifying imbalanced data in distance-based feature space," *Knowl. Inf. Syst.*, vol. 46, no. 3, pp. 707–730, 2016.
[46] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. Int. Conf. Mach. Learn.*, vol. 96. 1996, pp. 148–156.
[47] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[48] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[49] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.
[50] W. Zong, G.-B. Huang, and L. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, Feb. 2013.
[51] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Mach. Learn.*, vol. 31, no. 1, pp. 1–38, 2004.
[52] A. Lemmens and C. Croux, "Bagging and boosting classification trees to predict churn," *J. Marketing Res.*, vol. 43, no. 2, pp. 276–286, 2006.
[53] B. Zhu, B. Baesens, A. Backiel, and S. K. vanden Broucke, "Benchmarking sampling techniques for imbalance learning in churn prediction," *J. Oper. Res. Soc.*, vol. 69, no. 1, pp. 49–65, 2018.
[54] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, Mar. 2002.
[55] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[56] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft/Tech. Rep. MSR-TR-98-14, 1998.
[57] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.
[58] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[59] A. Jain, "Complete guide to parameter tuning in XGBoost (with codes in python)," Analytics Vidhya, 4, 2016.
[60] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.

POONGJIN CHO received the B.Sc. degree in industrial and management engineering, and mathematics from POSTECH, Pohang, South Korea, in 2013. He is currently pursuing the M.Sc. and Ph.D. degree in industrial engineering from Seoul National University, Seoul, South Korea. His research interests include pattern recognition, time-series analysis, econophysics, operations research, and the applications of these areas in crisis management, risk assessment, fraud detection, A.I. trading, demand forecasting, trend prediction, FinTech, and other topics related to data-driven analytics.

WOOJIN CHANG received the B.Sc. degree in fiber polymer engineering from Seoul National University, Seoul, South Korea, in 1997, and the M.Sc. degree in operations research and the Ph.D. degree in industrial engineering from the Georgia Institute of Technology, GA, USA, in 1998 and 2002, respectively. From 2002 to 2003, he was an Assistant Professor with the Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, NY, USA. He joined the Department of Industrial Engineering, Seoul National University, in 2004, where he is currently a Professor with the Financial Risk Engineering Laboratory.

JAE WOOK SONG received the B.Sc. degree in industrial engineering from the Georgia Institute of Technology, GA, USA, in 2010, and the Ph.D. degree in industrial engineering from Seoul National University, Seoul, South Korea, in 2016. From 2016 to 2018, he was a Data Scientist with the Big Data Analytics Group of Mobile Communications Business, Samsung Electronics, South Korea. Since 2018, he has been an Assistant Professor with the Department of Data Science, Sejong University, South Korea. His research interests include developing analytical frameworks for data-driven innovations in financial markets, investment decisions, and strategic management based on applications of complex systems, time-series analysis and forecasting, applied probability, and machine learning algorithms. He was a recipient of the Korean Operations Research and Management Science Society Best Dissertation Award, in 2016.