Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 869628, 15 pages
http://dx.doi.org/10.1155/2014/869628

Research Article

One-Step Dynamic Classifier Ensemble Model for Customer Value Segmentation with Missing Values

Jin Xiao,[1] Bing Zhu,[1] Geer Teng,[2] Changzheng He,[1] and Dunhu Liu[3]

[1] *Business School, Sichuan University, Chengdu 610064, China*
[2] *The Faculty of Social Development & Western China Development Studies, Sichuan University, Chengdu 610064, China*
[3] *Management Faculty, Chengdu University of Information Technology, Chengdu 610103, China*

Correspondence should be addressed to Dunhu Liu; 264885613@qq.com

Received 19 December 2013; Accepted 28 January 2014; Published 18 March 2014

Academic Editor: Chuangxia Huang

Scientific customer value segmentation (CVS) is the base of efficient customer relationship management, and customer credit scoring, fraud detection, and churn prediction all belong to CVS. In real CVS, the customer data usually include lots of missing values, which may affect the performance of CVS model greatly. This study proposes a one-step dynamic classifier ensemble model for missing values (ODCEM) model. On the one hand, ODCEM integrates the preprocess of missing values and the classification modeling into one step; on the other hand, it utilizes multiple classifiers ensemble technology in constructing the classification models. The empirical results in credit scoring dataset "German" from UCI and the real customer churn prediction dataset "China churn" show that the ODCEM outperforms four commonly used "two-step" models and the ensemble based model LMF and can provide better decision support for market managers.

1. Introduction

In the increasingly fierce market competition, the traditional resources in enterprises, such as product quality, price, and production capacity, have been unable to bring new competitiveness for the enterprises; customer relationship management (CRM) has become the new resource from which enterprises can gain benefit continuously. The core objective of CRM is to maximize the customer value for enterprises [1]. As we all know, each customer has different value for an enterprise; usually about 80% profits are created by 20% customers. The enterprise can get the most profits only if it devotes the limited resources to the most valuable customers, develops specialized customer strategy, and designs different products and services for different customers [2]. Therefore, scientific customer value segmentation (CVS) is the base for efficient CRM. In the last decades, CVS has been applied to many key business processes such as customer retention, customer growth analysis, and customer acquisition, and it provides important support for decision making in CRM [3, 4].

CVS is to classify customers according to their ability of creating value for enterprise, provide targeted products, services, and marketing models, enable enterprise to allocate resources more rationally, reduce cost effectively, and gain more profitable market penetration [5]. In fact, customer credit scoring, fraud detection, and churn prediction all belong to CVS, and their essential work is to classify customers according to different dimensions of customer value [6, 7].

The research methods of CVS can be roughly divided into three categories: (1) qualitative analysis (as for this method, the customer manager judges the value of any customer through selecting and reading the customer information. This method is highly subjective); (2) statistical analysis methods construct a statistical model for CVS (the representative methods include logistic regression [8], discriminant analysis [9], and so on); (3) machine learning methods (with the development of information technology, people began to utilize some models derived from machine learning in CVS in the 1990s, such as support vector machine [10], decision tree [11], artificial neural network [12], and multiple classifiers ensemble technology) [13].

With the in-depth study of the theory and practice of CVS, it is discovered that many customer data collected through questionnaires, interviews, and other means often include lots of missing values (MVs) [14, 15]. Yim et al. [15] have studied customer satisfaction through 450 questionnaires and found that 90 questionnaires contain a large number of MVs. The data not only from the questionnaires but also from the enterprises' CRM database often contain MVs. Take the enterprise Honeywell listed in Fortune Global 500 for example; although they have strict data collection standards, the missing rate of data in the customer database is still as high as 50% [14].

The MVs contained in CRM customer data influence the CVS effects to a large extent [16]. To solve this issue effectively, Kim et al. [17] have proposed a three-phase framework: (1) data collection and preprocess; (2) CVS modeling (it finds the main factors that affect customer value from customer information and constructs classification model to predict customer value based on these factors); (3) marketing strategies formulation (it proposes appropriate strategies for different customers to maximize their values for the enterprise according to the results obtained in Phase 2).

Scholars have done a lot of researches around the above framework. They usually preprocess the MVs to make the data complete and then establish CVS model. The two steps are carried out independently, so we call it "two-step" model. The simplest way of preprocess is listwise deletion (LD) [14], which deletes the instances with MVs from the dataset directly. For example, Subramania and Khare [18] have utilized pattern classification method in the diagnosis analysis of automotive warranty and service, and they adopt LD to handle MVs. This method is simple but it is easy to lose a lot of important information [19]. Therefore, the imputation methods are more popular. For instance, Lessmann and Voß [20] have proposed a support vector machine based hierarchical reference model for credit scoring and replaced MVs with the mean of the nonmissing values of the corresponding attribute before modeling; Paleologo et al. [21] have presented subagging model for credit scoring; they replace the missing values either by the maximum or by the minimum of the nonmissing values of the attribute; Li and Wang [22] have proposed Bayesian network technology based attribute fatigue analysis model in product development; they impute the MVs by EM method [23] before modeling.

The propositions mentioned above have made important contributions to customer value segmentation with MVs. However, some scholars have found that many CVS models are sensitive to data preprocess method, and the results are instable [24, 25]. In addition, as the most popular preprocess methods, imputation methods still have some disadvantages. The commonly used imputation methods are based on the assumption of random missing [26], so they all need to suppose that the data obey some distribution models. But in practice, a variety of missing mechanisms are often intertwined. If the assumption and the model are irrational, they are prone to data deflexion, which may lead to serious estimation bias and affect the learning effect of subsequent classifiers [27]. Therefore, the "two-step" CVS models need further improvement.

This study introduces multiple classifiers ensemble technology [28] to CVS and constructs one-step dynamic classifier ensemble model (ODCEM) for MVs, which integrates the preprocess of missing values and the classification modeling into one step. The empirical results in a customer credit scoring dataset and a customer churn prediction dataset show that the proposed method is superior to other models in CVS performance.

The structure of this study is organized as follows: it briefly introduces the commonly used processing methods for MVs in Section 2; proposes the work principle and detailed steps of ODCEM in Section 3; proceeds the experimental design and detailed results analysis in Section 4. Finally, the conclusions and future work are in Section 5.

2. The Commonly Used Processing Methods for MVs

In general, most customer value segmentation models, such as artificial neural network, logistic regression, and support vector machine, require that the customer data are complete. As long as one value of some attribute is missing, it cannot train the model [29]. At present, the main methods of handling MVs can be summarized into three categories: listwise deletion, imputation methods, and ensemble based methods.

2.1. Listwise Deletion. Listwise deletion (LD) [14] is the simplest method of handling MVs. For any instance with MVs in model training set, LD will delete it from the dataset directly. If the missing rate of the model training set (the ratio between the number of instances with MVs and the size of the dataset) is very small, LD will be effective, while, if the missing rate is large and the MVs are not distributed randomly, it may result in data deviation and gaining wrong conclusions [14]. At last, it cannot be used when each instance in model training set contains some MVs or the instances in test set also contain some MVs.

2.2. Imputation Methods. At present, the imputation methods are the most commonly used ones for dealing with MVs, in which each MV is replaced by a value generated by some mechanism according to the nonmissing values in the model training set. Many scholars have proposed a variety of MV imputation methods, among which some methods are very simple and applicable, such as mean substitution (MS) [30], K-nearest neighbours imputation (KI) [27], regression imputation (RI) [31], and EM imputation (EM) [23].

Mean substitution (MS) [30] usually divides the attributes into nominal and numeric while dealing with MVs. For nominal attribute, the MV is replaced by the most common attribute value, while for numerical one, the MV is replaced by the average of all nonmissing values of the corresponding attribute.

K-nearest neighbours' imputation [27] uses an instance based algorithm to impute the MVs. Every time it finds a MV in a current instance, it computes the K-nearest neighbours of the instance and imputes a value from them. For nominal value, the most common value among all neighbours is taken, and for numerical value we will use the average value.

RI method [31] needs to select some independent attributes for predicting the MV first and then constructs the regression equation to estimate the MV, that is, replaces the MV with its conditional expectation. In detail, given a MV for an attribute x, suppose that q attributes have been observed for that instance. The records where these $q + 1$ attributes are available define a training set, and a regression model to predict x from the q predictors is fitted. Finally, the fitted model provides a prediction for the initial MV of x.

EM imputation [23] finds the maximum likelihood estimates recursively according to the observed data and consists of two steps: expectation step (E-step) and maximization step (M-step). In E-step, it calculates the expectation of the complete data sufficient statistics given the observed data and current parameter estimates and updates the parameter estimates through the maximum likelihood approach based on the current values of the complete sufficient statistics in M-step. It repeats the two steps till the parameter estimation converges, and the expectation of each MV in the final E-step is regarded as the imputation value.

In fact, all the above imputation methods belong to single imputation. Single imputation replaces each MV with one value, which cannot reflect the uncertainty of MVs well. Thus, Rubin [32] has proposed multiple imputation (MI) method. In this method, each MV is imputed m times by the same imputation algorithm, which uses a model that incorporates some randomness. As a result, m "complete" datasets are generated, and usually the average of the estimates across the samples is used to generate the final imputation value. The main disadvantage of MI imputation is the large calculating cost.

2.3. Ensemble Based Methods. Recently, some scholars have tried to utilize multiple classifiers ensemble technology to construct the classification model for MVs. For example, Krause and Polikar [33] and Mohammed et al. [34] have proposed Learn^{++} method for missing features (abbreviated as LMF) which classifies the data with MVs directly. It selects some attribute subsets in the whole attribute space, obtains a number of training subsets by mapping, and then trains a base classifier in each training subset. For each test instance x^* (may contain MVs), LMF finds the base classifiers which can classify it and combines the classification results of the selected classifiers by voting to get its final classification result. The empirical results show that LMF method can achieve better classification performance. For more detailed process of Learn^{++}, please refer to [34].

In theory, LMF also belongs to one-step ensemble strategy. However, Mohammed et al. have also pointed out that LMF cannot classify the test instance with many MVs because we cannot find any available base classifier for it [34]. At last, for each test instance x^*, LMF method gets the final classification results through combining the results of all available base classifiers. However, redundancy may exist among the base classifiers. Therefore, it is expected to improve the classification performance if an appropriate classifier subset can be selected to ensemble.

3. One-Step Dynamic Ensemble Model for Missing Values

3.1. Basic Idea. In fact, there are many data issues such as noise and imbalanced class distribution except missing values in customer value segmentation (CVS). In this study, we mainly focus on the issue of CVS with missing values and propose one-step dynamic classifier ensemble model (ODCEM) for missing values, while for the other issues such as imbalanced class distribution that may exist in the CVS dataset, we need to preprocess them first and then construct ODCEM model.

The terms of one-step ensemble in ODCEM model contain two meanings: first, it integrates the preprocess of missing values in Phase 1 with the customer classification modeling in Phase 2 from the "three-phase" CVS framework proposed by Kim et al. [17] and reduces the dependence of assumption about missing mechanism and data distribution model; second, it introduces multiple classifiers ensemble technique to customer value classification modeling.

Suppose a CVS issue contains n attributes; its training set D_{train} and test set D_{test} contain m_1 and m_2 customer instances, respectively. In addition, all the instances can be divided into L classes according to their values for the enterprise, and both D_{train} and D_{test} contain some MVs.

ODCEM model mainly includes two phases: training base classifiers and classifying test instances. In training phase, it first divides D_{train} into three subsets according to the missing rate of instances: D_1, D_2, and D_3 (the reason of the number of subsets being three is that the three subsets correspond to low missing level set, middle missing level set, and high missing level set, resp.), and then in each subset it assigns different sampled weights to the attributes according to different numbers of MVs in each attribute, selects T attribute subsets randomly, obtains a series of training subsets by mapping like the random subspace method [35], deletes the instances with MVs in each training subset, and then trains the classification model to compose the base classifier pool (BCP). In classifying test instances phase, for each test instance $x_j^* \in D_{\text{test}}$ $(j = 1, 2, \ldots, m_2)$, it finds K-nearest neighbours of x_j^* from D_{train} to compose the local area L_a of x_j^*, selects some classifiers with best classification performance in L_a from BCP to classify x_j^*, and finally obtains the final classification result of x_j^* by weighted voting. The process of ODCEM model is described in Figure 1.

It is notable that, in ODCEM model, if there are only a few MVs in D_{train}, most of the training instances will be assigned to the low missing level subset D_1, and we can get the base classifiers with good enough classification performance by sufficient training instances, which can ensure satisfactory CVS effect of ODCEM, while if there are lots of MVs in D_{train} and the subsets D_1, D_2, and D_3 may all contain a certain number of instances, it can also find the base classifiers in subset D_1 or D_2 containing fewer MVs to classify the test instance x_j^*. In short, it can always find some base

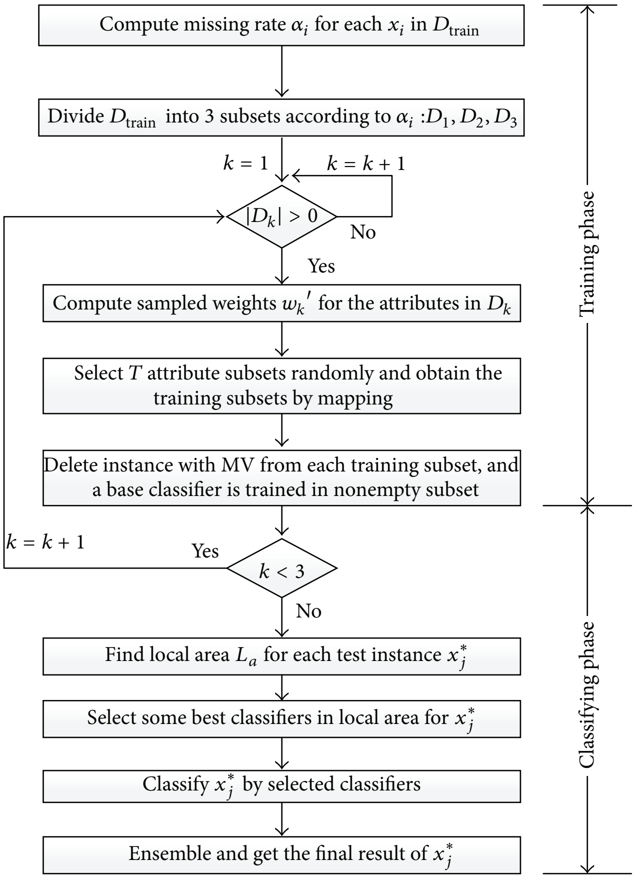

Compute missing rate α_i for each x_i in D_{train}
Divide D_{train} into 3 subsets according to α_i :D_1, D_2, D_3
$k = 1$
$k = k + 1$
$|D_k| > 0$
No
Yes
Compute sampled weights $w_k{}'$ for the attributes in D_k
Select T attribute subsets randomly and obtain the training subsets by mapping
Delete instance with MV from each training subset, and a base classifier is trained in nonempty subset
$k = k + 1$
Yes
$k < 3$
No
Find local area L_a for each test instance x_j^*
Select some best classifiers in local area for x_j^*
Classify x_j^* by selected classifiers
Ensemble and get the final result of x_j^*
Training phase
Classifying phase

Figure 1: Flow chart of ODCEM algorithm.

classifiers for a given test instance. Thus, ODCEM method can make up for the disadvantage of LMF method proposed by Mohammed et al. [34] to a large extent.

In the following content, we will describe the process of ODCEM model in detail.

3.2. Train Base Classifiers. To train the base classifiers, ODCEM first divides D_{train} into 3 subsets according to the missing rate of the instances. For each instance $x_i \in D_{\text{train}}$, its missing rate α_i is defined as follows:

$$\alpha_i = \frac{n_{\text{miss}}}{n}, \quad i = 1, 2, ..., m_1, \tag{1}$$

where n_{miss} is the number of missing values in instance x_i and n is the total number of attributes. It is easy to know that $0 \leq \alpha_i \leq 1$.

After getting the value of α_i $(i = 1, 2, ..., m_1)$, we rank all training instances according to the order of α_i from small to large. Thus, the more frontier the instance is, the fewer missing values the instance has; even there may be no MV. Further, we divide the range of α_i into 3 intervals: $[0, 1/3)$, $[1/3, 2/3)$, and $[2/3, 1]$ and divide the whole training set D_{train} into 3 subsets according to the intervals: D_1, D_2, and D_3. Therefore, D_1 contains the instances with the missing rate $\alpha_i \in [0, 1/3)$, and similarly, D_3 contains the instances with the missing rate $\alpha_i \in [2/3, 1]$.

In order to train base classifiers in each subset D_k $(k = 1, 2, 3)$, we select a series of attribute subsets randomly according to the basic idea of random subspace (RSS) [35]. As the number of MVs in different attributes is often different, the less the number is, the more information the attribute contains and the larger the possibility to be selected is. As for subset D_k, if it is nonempty, then the sampled weight w_{kj} of attribute f_j $(j = 1, 2, \ldots, n)$ in D_k is calculated as follows:

$$w_{kj} = \frac{|D_k|}{p_{kj}}, \quad j = 1, 2, \ldots, n;\ k = 1, 2, 3, \tag{2}$$

where p_{kj} is the number of MVs in the column of attribute f_j in subset D_k and $|D_k|$ is the total number of instances in D_k. Especially, if $p_{kj} = 0$, that is, there is no MV in the column of attribute f_j, we let $w_{kj} = 2$ directly; namely, assign much larger sampled weight for the attribute f_j compared with the attribute with MVs. Finally, the weight vector $w_k = (w_{k1}, w_{k2}, \ldots, w_{kn})$ is normalized, and the final sampled weight of attribute f_j in D_k is obtained:

$$w'_{kj} = \frac{w_j}{\sum w_{kj}}, \quad j = 1, 2, \ldots, n. \tag{3}$$

Select T attribute subsets randomly according to the sampled weight vector $w'_k = (w'_{k1}, w'_{k2}, \ldots, w'_{kn})$; the number of attributes in each attribute subset is equal to half of the total attributes [35]; then obtain some training subsets by mapping. In fact, the training subsets obtained at this moment may contain MVs, while the commonly used classification models such as neural network, support vector machine, and logistic regression require that the training set cannot contain MVs. Thus, we delete the instance with MV from the training subset, and if the remaining training subset is nonempty, then train a base classifier by it. Finally, all the trained base classifiers consist of a base classifier pool (BCP). It is worth noting that the number of base classifiers in BCP varies in different datasets, which may be affected by missing rate of the dataset, the size of the dataset, and so forth.

3.3. Classify the Test Instances. In the above section, a series of base classifiers are trained in the training set. In this section, we will classify all the test instances in test set by the trained base classifiers. The ODCEM model proposed in this study belongs to dynamic classifier ensemble selection [36], in which a classifier subset is selected out from BCP for each test instance $x_j^* \in D_{\text{test}}$ $(j = 1, 2, \ldots, m_2)$.

To achieve this process, it needs to select K-nearest neighbours from the total training set D_{train} to compose the local area (called L_a) of x_j^*. In this study, we choose Euclidean distance measure to calculate the distance between x_j^* and all instances in D_{train}. However, it cannot be calculated sometimes because there is MV in x_j^* or in the training instance. For test instance $x_j^* = (x_{j,1}^*, x_{j,2}^*, \ldots, x_{j,n}^*)$ and any instance $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$ in D_{train}, we define two MV indication vectors: $u_j = (u_{j,1}, u_{j,2}, \ldots, u_{j,n})$ and $v_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,n})$ as follows:

$$u_{jk} = \begin{cases} 0 & \text{if the value of } x^*_{j,k} \text{ is missing} \\ 1 & \text{else,} \end{cases}$$
$$v_{ip} = \begin{cases} 0 & \text{if the value of } x_{i,p} \text{ is missing} \\ 1 & \text{else,} \end{cases} \tag{4}$$

where $k, p = 1, 2, \ldots, n$. Further, we suppose only $s\,(s \le n)$ attribute values are not missing in test instance x^*_j; that is, $u_{je_1} = u_{je_2} = \cdots = u_{je_s} = 1$; here, $1 \le e_t \le n$, $t = 1, 2, \ldots, s$. Thus, the distance between the test sample x^*_j and the instance x_i in D_{train} can be calculated as follows:

$$d_{j,i} = \begin{cases} \sqrt{\sum_{t=1}^{s} \left(x^*_{j,e_t} - x_{i,e_t} \right)^2}, & \text{if } v_{i,e_1} = v_{i,e_2} = \cdots = v_{i,e_s} = 1 \\ \infty, & \text{else.} \end{cases} \tag{5}$$

In (5), when the distance between two instances cannot be calculated, we take it as ∞.

After finding the local area L_a of x^*_j through (5), we select some suitable classifiers to classify K instances in L_a (a classifier is selected if there is no missing value in the corresponding columns in L_a with all attributes in the feature subset used for training this classifier), calculate the classification accuracy of each classifier in L_a, and select half number (denoted by N) of the selected base classifiers with higher classification accuracy, and their classification accuracy in L_a is $\text{acc}_1, \text{acc}_2, \ldots, \text{acc}_N$, respectively. Classify the test instance x^*_j with N selected classifiers; for each base classifier C_r $(r = 1, 2, \ldots, N)$, it will output a probability estimation $P(f(x^*_j) = c \mid C_e)$, which means the probability of x^*_j belongs to the category of c $(c = 1, 2, \ldots, L)$. The final prediction result of test instance x^*_j is obtained through weighted voting:

$$\rho = \arg_c \left(\max_{c=1,2,\ldots,L} \left(\frac{1}{N} \sum_{r=1}^{N} z_r P\left(f\left(x^*_j \right) = c \mid C_r \right) \right) \right), \tag{6}$$

where $z_r = \text{acc}_r / \sum \text{acc}_r$, $r = 1, 2, \ldots, N$.

3.4. Pseudocode of Model. The pseudocode of ODCEM model is as follows.

Input. Training set D_{train}, test set D_{train}, number of nearest neighbours K, and the number of attribute subsets T are selected in each subset D_k $(k = 1, 2, 3)$.

Output. The final classification result of each instance x^*_j in D_{test} is as follows.

```
(1) Calculate the missing rate $\alpha_i$ of each instance in $D_{\text{train}}$
    according to (1) and divide the entire training set into
    3 subsets according to $\alpha_i$: $D_1$, $D_2$, and $D_3$.
(2) For each subset $D_k$ $(k = 1, 2, 3)$, if $D_k$ is nonempty,
    then
    (2.1) calculate the sampled weights $w'_k = (w'_{k1}, w'_{k2}, \ldots, w'_{kn})$
          of all attributes according to (3);
    (2.2) select $T$ attribute subsets randomly from
          attribute space according to $w'_k$ and obtain $T$
          training subsets $S_i$ $(i = 1, 2, \ldots, T)$ by mapping;
    (2.3) delete the instance with MV in $S_i$, and if the
          remaining $S_i$ is nonempty, then a base classifier
          is trained with $S_i$ and added to BCP.
(3) For each test instance $x^*_j \in D_{\text{test}}$,
    (3.1) find $K$-nearest neighbours of $x^*_j$ from the entire
          training set $D_{\text{train}}$ to compose its local area $L_a$
          according to (5);
    (3.2) classify the instances in $L_a$ with all base classifiers
          and select half of the base classifiers in BCP
          with the highest classification accuracy in $L_a$;
    (3.3) classify $x^*_j$ with the selected base classifiers
          and obtain the final classification result of $x^*_j$
          according to (6).
```

4. The Empirical Study

In order to analyze the CVS performance of ODCEM proposed in this study, we conducted experiments in the credit scoring dataset "German" from UCI [37] and credit card customer churn prediction dataset "China churn" of one commercial bank in Sichuan province China. At the same time, we compared the CVS performance of ODCEM with that of four commonly used "two-step" models for MVs, which impute MVs by K-nearest neighbours imputation (KI), mean substitution (MS), EM imputation (EM), and regression imputation (RI), respectively, and then constructed multiple classifier systems with subagging method [21]. It is worth noting that listwise deletion (LD) imputation was not referred to in our experiments because of its obvious disadvantage, and multiple imputation (MI) method was not selected as the benchmark for its high computation cost. Finally, we also compared one-step ensemble selection strategy ODCEM with the ensemble based method LMF proposed by Mohammed et al. [34].

4.1. Description of the Datasets. The first dataset used in the study is "German," a credit card customer credit scoring dataset from German [37]. There are 20 attributes and 1 class label C in the dataset, in which 7 attributes are numeric and 13 attributes are qualitative. The class label includes two different states $\{good, bad\}$ which divide the customers into two classes: good credit and bad credit. There are 1000 customer instances in the dataset: 700 customers with good credit and 300 customers with bad credit. In addition, there is no MV in it.

The second dataset is about churn prediction of credit card customer from one commercial bank in Sichuan province, China ("China churn"). The data interval is 2010.5–2010.12. According to the basic principle of churn prediction attribute selection and considering the availability of data,

Table 1: Attribute description of “China churn” dataset.

Attribute	Name	Missing rate (%)
x_1	Total consumption times	7.37
x_2	Total consumption amount	6.76
x_3	Total cash times	4.95
x_4	Customer survival time	8.63
x_5	Total contributions	11.12
x_6	Valid survival time	8.05
x_7	Average amount ratio	10.75
x_8	Whether associated charge	12.18
x_9	Consumption times in the last 1 month	2.09
x_{10}	Consumption times in the last 2 months	3.13
x_{11}	Consumption times in the last 3 months	8.08
x_{12}	Consumption times in the last 4 months	9.31
x_{13}	Consumption times in the last 5 months	3.87
x_{14}	Cash times in the last 1 month	5.07
x_{15}	Cash times in the last 2 months	7.04
x_{16}	Cash times in the last 3 months	8.97
x_{17}	Cash times in the last 6 months	10.60
x_{18}	Months of transaction times reducing continuously	8.48
x_{19}	If overdue in the last 1 month	7.80
x_{20}	If overdue in the last 2 months	13.76
x_{21}	Amount usage ratio in the last 1 month/historical average usage ratio	6.11
x_{22}	Sex	2.70
x_{23}	Annual income	14.84
x_{24}	Nature of work industry	11.18
x_{25}	Education	8.63

we selected 25 prediction attributes (see Table 1). For the customer class label, we defined the churn customer as someone who canceled card from May 2010 to October 2010 or did not consume for 3 months. After simple data cleaning, we obtained 3255 customer instances from the database finally, in which there are 302 churn customers and 2953 nonchurn customers. The churn rate is 9.28% and it belongs to class imbalanced dataset. Meanwhile, it includes a lot of MVs, and the missing rate of each attribute is in Table 1.

4.2. Experimental Design. As there is no MV in “German” dataset, we generated the MVs artificially and analyzed the CVS performance of various models under different missing condition. Depending on the reason why MVs have been produced, the missing mechanism can be classified into three types [19]: MCAR (missing completely at random), MAR (missing at random), and MNAR (missing not at random). Suppose that a dataset D contains two parts: D_{obs} and D_{mis}, where D_{obs} stands for all observed data and D_{mis} stands for all MVs. If $P(R \mid D_{\text{obs}}, D_{\text{mis}}) = P(R \mid D_{\text{obs}})$, where R is the event we concerned, then we call it MAR, and if $P(R \mid D_{\text{obs}}, D_{\text{mis}}) = P(R)$, then we call it MCAR, while if it does not meet the above two missing mechanisms, we call it MNAR. In this study, we considered the above three mechanisms in “German” dataset, and let the missing level $\theta = 5\%, 10\%, 20\%, 30\%$, and 40%, respectively. In order to facilitate this process, we adopted the method proposed by Fayyad and Irani [38] to prediscretize the continuous attributes, and then three kinds of MVs were produced as follows.

(1) *MCAR*. For each instance, a random number $r \in (0, 1)$ was generated. If $r < \theta$, then ceil $(20 * r)$ attributes were selected randomly (here ceil() is an up-rounding function in Matlab and 20 is the number of attributes in “German” dataset), and let the values of these attributes be missing. Since the MVs of each instance are random, it should belong to missing completely at random.

(2) *MAR*. For each instance, a random number $r \in (0, 1)$ was generated, and if $r < \theta$, then two attributes x_i, x_j, $(i < j)$ were selected randomly. If $x_i = 1$ (after prediscretization in “German” dataset, the smallest value of each attribute is just 1, and thus it can make this condition be suitable for any attribute), let the value of x_j be missing. The MVs of x_j are only related to the value of attribute x_i and have nothing to do with its own values. Therefore, it should belong to missing at random.

(3) *MNAR*. For each instance, a random number $r \in (0, 1)$ was generated, and if $r < \theta$, then an attribute x_i was selected randomly. If $x_i = 1$, let the value of x_i be missing. The MVs of x_i are related to its own values, which should belong to missing not at random.

In this study we chose support vector machine (SVM) to generate the base classifier for its popularity and immense success in various CVS tasks [18]. Thus, the four "two-step" CVS models are abbreviated as KI-SVM, MS-SVM, EM-SVM, and RI-SVM. When training SVM, the choice of kernel function is very important. We found that radial basis kernel (RBK) based classifier could obtain the best performance through experimental comparison, so we chose it as the kernel function of SVM. Meanwhile, there are two important parameters in ODCEM model: the number of nearest neighbours K in the local area and the number of attribute subset T selected in each subset D_k ($k = 1, 2, 3$). We conducted the sensitivity analysis of the two parameters and experimented in two datasets with seven different values of K: 3, 5, 7, 9, 11, 13, and 15 and seven different values of T: 10, 15, 20, 25, 30, 35, and 40, respectively. We found that the performance of ODCEM with different values of K showed some fluctuations, and it performed best when $K = 5$. As for the parameter T, it is found that the performance of ODCEM rose first with the increase of T and achieved the best when T equaled about 30 and then showed a little fluctuation when $T > 30$. The detailed analysis is omitted because of the space consideration, and we made $K = 5$, $T = 30$. Further, as for the subagging ensemble strategy used in four "two-step" models, Paleologo et al. [21] have found that the number of base classifiers in between 20 and 50 is reasonable, and thus we let it be the maximum 50 in four "two-step" models as well as the LMF model.

Further, the class distribution of both datasets is imbalanced. If we train CVS model with such data directly, then it tends to classify all customers as the majority class. There are many techniques to deal with class imbalance data, including oversampling and downsampling. In this study, what we concern most is not the optimal match between the six models referred and the methods handling class imbalance data, but the CVS performance of six models in the condition of MVs. Therefore, without loss of generality, we adopted oversampling method to balance the class distribution of training set and then conducted the experiments by the six models. In addition, to compare the performance of ODCEM proposed in this study with the other five models, we adopted 10-fold cross-validation (CV10) which divides the entire dataset into 10 equal parts randomly and takes one part as test set and the other nine parts as training set every time; the rotation of 10 times is called CV10.

As for the four "two-step" models, we imputed the MVs with SPSS 17.0 for EM-SVM and RI-SVM first and then conducted model training and classification on the platform of Matlab 2008b, while the other two simpler "two-step" models KI-SVM and MS-SVM, as well as ODCEM and ensemble based LMF, were implemented on the platform of Matlab 2008b directly. Finally, all experiments were performed on a dual-processor 3.0 GHz Pentium 4 Windows computer with 2 GB RAM, and the final classification result was the average of 10 times CV10 in each case.

4.3. Evaluation Criteria. To evaluate the performance of the models referred to in this study, we introduced the confusion matrix in Table 2. On this basis, four commonly used evaluation criteria were adopted [39]:

(1) total accuracy = ((TP + TN)/(TP + FN + FP + TN)) × 100%;

(2) the area under the receiver operating characteristic curve (AUC). The receiver operating characteristic (ROC) curve is an important evaluation criterion of classification model in the data with imbalanced class distribution [40]. For an issue of two classes, the ROC graph is a true positive rate-false positive rate graph, where y-axis is true positive rate (TP/(TP + FN) × 100%) and the x-axis is false positive rate (FP/(FP + TN) × 100%). However, sometimes it is difficult to compare ROC curves of different models directly, so AUC is more convenient and popular;

(3) type I accuracy = (TP/(TP + FN)) × 100%;

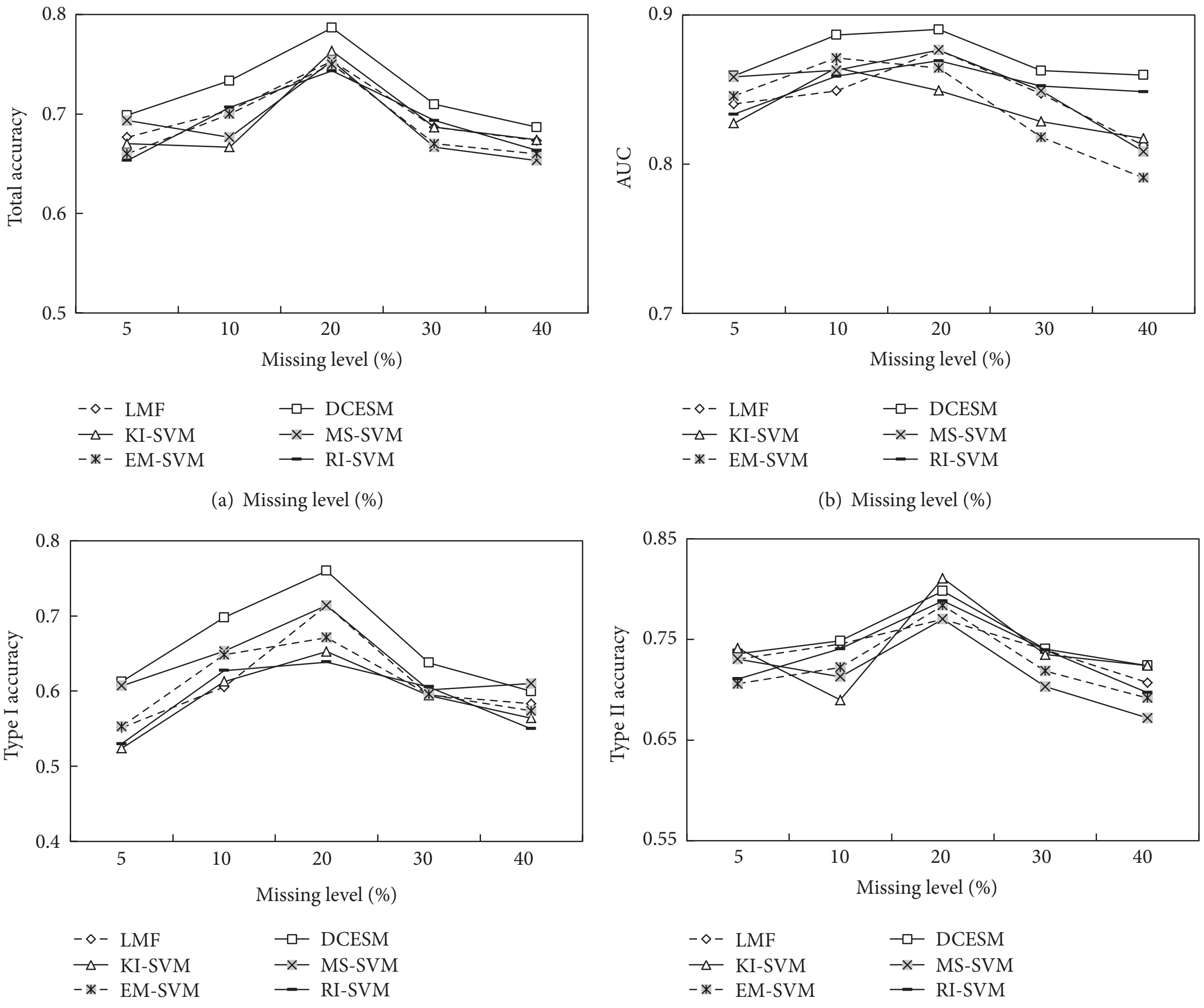
(4) type II accuracy = (TN/(FP + TN)) × 100%.

4.4. Experimental Results Analysis in "German" Dataset. Table 3 shows the credit scoring performance of six models in "German" dataset with MCAR type MVs. The results show that, (1) for the total accuracy, AUC, and type I accuracy, ODCEM outperforms the other five models under each missing level; (2) for the type II accuracy, ODCEM outperforms the other five models when $\theta = 10\%$ and 30%, and it performs more poorly than KI-SVM when $\theta = 5\%$, 20%, and 40%. Compared with the type II accuracy, we are usually concerned more about the AUC and type I accuracy of a model in class imbalanced CVS issue. Thus, we can conclude that the overall credit scoring performance of ODCEM is better than that of the other models referred to in this study in "German" dataset with MCAR type MVs.

In addition, the performance rank of six models on each evaluation criterion in five missing levels, respectively, is also shown in Table 3, and the last row is the average rank, which can be regarded as a criterion of the overall performance of the models. Therefore, we can rank the six models according to the overall performance for MCAR type MVs from high to low as follows: ODCEM, LMF, MS-SVM, KI-SVM, RI-SVM, and EM-SVM. It is notable that LMF outperforms the four "two-step" models, and EM-SVM performs the most poorly in this case.

Figure 2 shows the trend of credit scoring performance of six models in "German" dataset with MCAR type MVs. It can be seen from the figure that the performance of each model does not decline obviously with the increase of missing level but appears with great fluctuation, and the general trend is increasing first and then reducing. This may be related to the production means of MVs of MCAR in this study, or related to the inherent characteristics of "German" dataset.

Similarly, the performance rank of six models on each evaluation criterion in five missing levels is also shown in Table 4. Thus, we can rank the six models according to the overall performance for MAR type MVs from high to low as follows: ODCEM, EM-SVM, RI-SVM, MS-SVM, KI-SVM, and LMF. It is notable that the performance of EM-SVM is



(a) Missing level (%)

(b) Missing level (%)

(c) Missing level (%)

(d) Missing level (%)

Figure 2: The trend of classification performance in “German” dataset with MCAR type MVs.

Table 2: Confusion matrix.

	Predicted positive	Predicted negative
Actual positive (minority class)	TP (the number of true positives)	FN (the number of false negatives)
Actual negative (majority class)	FP (the number of false positives)	TN (the number of true negatives)

only poorer than that of ODCEM, and the performance of LMF is the poorest.

The trend of customer credit scoring performance of six models in “German” dataset with MAR type MVs is shown in Figure 3. It is shown that the performance of six models reduces quickly with the increase of missing level. In particular, when the missing level is low, such as $\theta = 5\%$, the ensemble based model LMF can achieve good classification performance, which is only poorer than that of ODCEM and EM-SVM, while with the increase of missing level, such as when $\theta \geq 20\%$, the performance of LMF is poorer than that of the other five models. The results also demonstrate that ODCEM model proposed in this study can overcome the disadvantages of LMF model to a large extent and can achieve better performance in high missing level than LMF.

The performance of six models in “German” dataset with MNAR type MVs is shown in Table 5. Although ODCEM only achieves comparable performance with MS-SVM and EM-SVM when $\theta = 5\%$, the performance of ODCEM is better than that of the other models when $\theta = 10\%$, 20%, 30%, and 40%. Thus, we can conclude that the overall credit scoring performance of ODCEM is still better than that of the other five models with MNAR type MVs in “German” dataset. Further, according to the average performance rank in the last row of Table 5, the six models can be ranked as follows: ODCEM, KI-SVM, MS-SVM, EM-SVM, LMF, and RI-SVM.

Table 3: Comparison of performance in "German" dataset with MCAR type MVs.

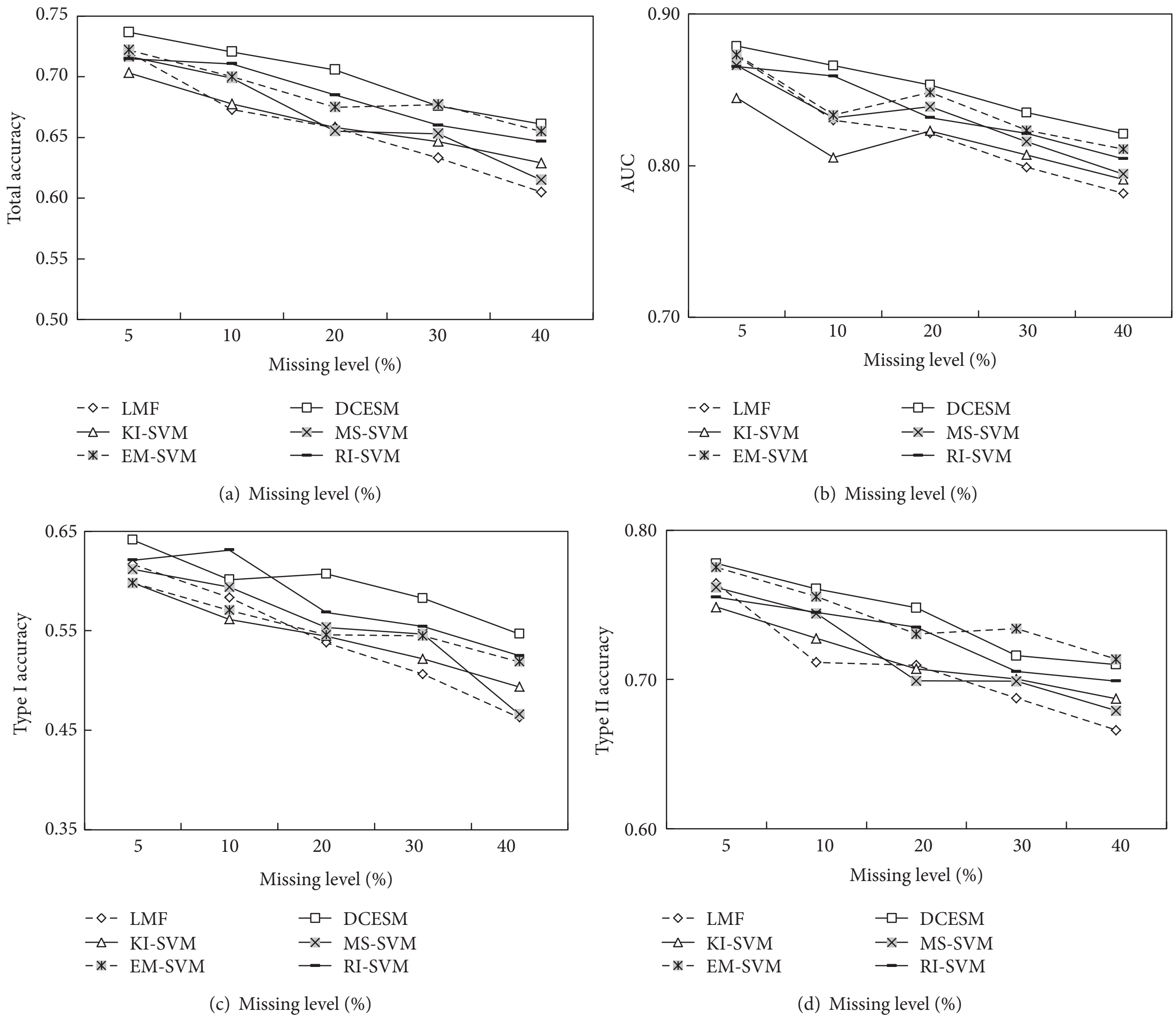
Missing level	Evaluation criteria	LMF	ODCEM	KI-SVM	MS-SVM	EM-SVM	RI-SVM
θ = 5%	Total accuracy	0.6767 (3)	**0.6987** (1)	0.6700 (4)	0.6933 (2)	0.6600 (5)	0.6533 (6)
	AUC	0.8402 (4)	**0.8592** (1)	0.8274 (6)	0.8583 (2)	0.8456 (3)	0.8334 (5)
	Type I accuracy	0.5513 (4)	**0.6124** (1)	0.5239 (6)	0.6069 (2)	0.5529 (3)	0.5298 (5)
	Type II accuracy	0.7304 (3.5)	0.7357 (2)	**0.7412** (1)	0.7304 (3.5)	0.7059 (6)	0.7108 (5)
θ = 10%	Total accuracy	0.7033 (3)	**0.7333** (1)	0.6667 (6)	0.6767 (5)	0.7000 (4)	0.7067 (2)
	AUC	0.8490 (6)	**0.8865** (1)	0.8645 (3)	0.8628 (4)	0.8710 (2)	0.8589 (5)
	Type I accuracy	0.6052 (6)	**0.6979** (1)	0.6127 (5)	0.6529 (2)	0.6481 (3)	0.6272 (4)
	Type II accuracy	0.7454 (2)	**0.7485** (1)	0.6898 (6)	0.7130 (5)	0.7222 (4)	0.7407 (3)
θ = 20%	Total accuracy	0.7533 (3.5)	**0.7867** (1)	0.7633 (2)	0.7533 (3.5)	0.7500 (5)	0.7433 (6)
	AUC	0.8764 (2.5)	**0.8902** (1)	0.8493 (6)	0.8764 (2.5)	0.8643 (5)	0.8692 (4)
	Type I accuracy	0.7138 (2.5)	**0.7598** (1)	0.6526 (5)	0.7138 (2.5)	0.6712 (4)	0.6384 (6)
	Type II accuracy	0.7703 (5.5)	0.7982 (2)	**0.8108** (1)	0.7703 (5.5)	0.7838 (4)	0.7883 (3)
θ = 30%	Total accuracy	0.6867 (3.5)	**0.7097** (1)	0.6867 (3.5)	0.6667 (6)	0.6700 (5)	0.6933 (2)
	AUC	0.8472 (4)	**0.8625** (1)	0.8286 (5)	0.8491 (3)	0.8180 (6)	0.8524 (2)
	Type I accuracy	0.5946 (5)	**0.6375** (1)	0.5941 (6)	0.6016 (3)	0.5956 (4)	0.6054 (2)
	Type II accuracy	0.7396 (2.5)	**0.7406** (1)	0.7349 (4)	0.7031 (6)	0.7188 (5)	0.7396 (2.5)
θ = 40%	Total accuracy	0.6733 (3)	**0.6867** (1)	0.6740 (2)	0.6533 (6)	0.6600 (5)	0.6633 (4)
	AUC	0.8172 (3)	**0.8596** (1)	0.8484 (2)	0.8083 (5)	0.7908 (6)	0.8125 (4)
	Type I accuracy	0.5832 (3)	**0.6101** (1)	0.5639 (5)	0.5997 (2)	0.5736 (4)	0.5498 (6)
	Type II accuracy	0.7071 (3)	**0.7240** (1.5)	**0.7240** (1.5)	0.6719 (6)	0.6919 (5)	0.6971 (4)
Average rank		3.63	1.13	4.00	3.83	4.40	4.03

Note: the bold-face in Table 3 shows the maximum of each row. The numbers in parentheses are the ranks of the six models with the corresponding evaluation criterion in each row.

Table 4: Comparison of performance in "German" dataset with MAR type MVs.

Missing level	Evaluation criteria	LMF	ODCEM	KI-SVM	MS-SVM	EM-SVM	RI-SVM
θ = 5%	Total accuracy	0.7200 (3)	**0.7367** (1)	0.7033 (6)	0.7167 (4)	0.7220 (2)	0.7148 (5)
	AUC	0.8724 (3)	**0.8789** (1)	0.8447 (6)	0.8662 (4)	0.8732 (2)	0.8653 (5)
	Type I accuracy	0.6167 (3)	**0.6411** (1)	0.5979 (6)	0.6117 (4)	0.5980 (5)	0.6209 (2)
	Type II accuracy	0.7643 (3)	**0.7777** (1)	0.7485 (6)	0.7617 (4)	0.7752 (2)	0.7550 (5)
θ = 10%	Total accuracy	0.6730 (6)	**0.7205** (1)	0.6777 (5)	0.6990 (4)	0.7000 (3)	0.7107 (2)
	AUC	0.8300 (5)	**0.8661** (1)	0.8054 (6)	0.8318 (4)	0.8334 (3)	0.8592 (2)
	Type I accuracy	0.5833 (4)	0.6012 (2)	0.5615 (6)	0.5938 (3)	0.5705 (5)	**0.6310** (1)
	Type II accuracy	0.7114 (6)	**0.7606** (1)	0.7275 (5)	0.7441 (4)	0.7555 (2)	0.7448 (3)
θ = 20%	Total accuracy	0.6580 (5)	**0.7057** (1)	0.6583 (4)	0.6553 (6)	0.6750 (3)	0.6850 (2)
	AUC	0.8215 (6)	**0.8533** (1)	0.8230 (5)	0.8390 (3)	0.8484 (2)	0.8317 (4)
	Type I accuracy	0.5380 (6)	**0.6070** (1)	0.5446 (5)	0.5534 (3)	0.5459 (4)	0.5683 (2)
	Type II accuracy	0.7094 (4)	**0.7480** (1)	0.7071 (5)	0.6990 (6)	0.7303 (3)	0.7350 (2)
θ = 30%	Total accuracy	0.6331 (6)	0.6759 (2)	0.6467 (5)	0.6531 (4)	**0.6772** (1)	0.6600 (3)
	AUC	0.7990 (6)	**0.8350** (1)	0.8071 (5)	0.8159 (4)	0.8233 (2)	0.8213 (3)
	Type I accuracy	0.5063 (6)	**0.5826** (1)	0.5217 (5)	0.5467 (3)	0.5447 (4)	0.5542 (2)
	Type II accuracy	0.6874 (6)	0.7159 (2)	0.7002 (4)	0.6987 (5)	**0.7340** (1)	0.7053 (3)
θ = 40%	Total accuracy	0.6050 (6)	**0.6610** (1)	0.6290 (4)	0.6150 (5)	0.6550 (2)	0.6467 (3)
	AUC	0.7818 (6)	**0.8211** (1)	0.7910 (5)	0.7945 (4)	0.8109 (2)	0.8046 (3)
	Type I accuracy	0.4627 (6)	**0.5466** (1)	0.4934 (4)	0.4658 (5)	0.5187 (3)	0.5249 (2)
	Type II accuracy	0.6660 (6)	0.7100 (2)	0.6871 (4)	0.6789 (5)	**0.7134** (1)	0.6989 (3)
Average rank		5.10	1.20	5.05	4.20	2.60	2.85

Note: the bold-face in Table 4 shows the maximum of each row. The numbers in parentheses are the ranks of the six models with the corresponding evaluation criterion in each row.



(a) Missing level (%)

(b) Missing level (%)

(c) Missing level (%)

(d) Missing level (%)

Figure 3: The trend of classification performance in “German” dataset with MAR type MVs.

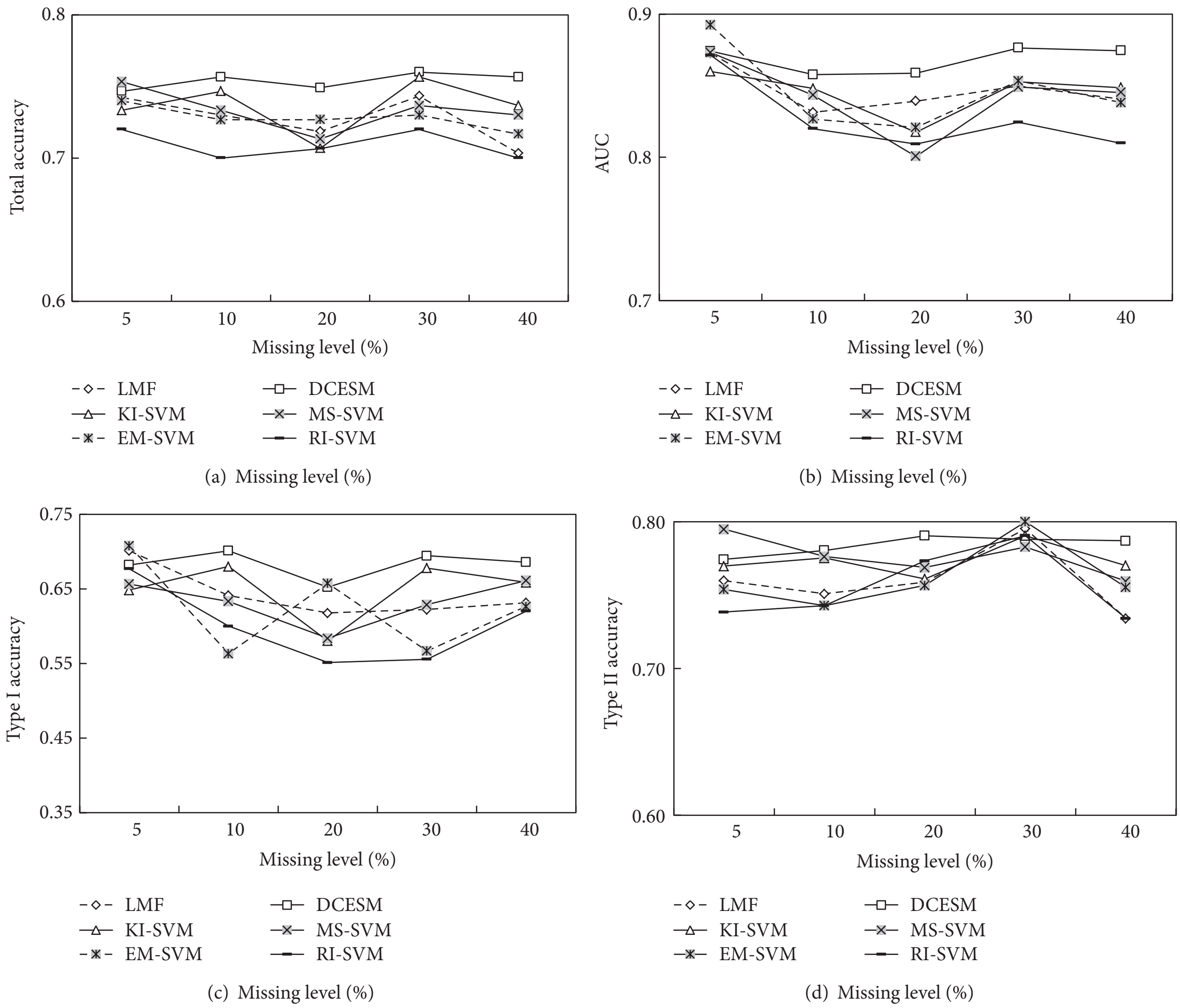
The trend of credit scoring performance of six models in “German” dataset with MNAR type MVs is shown in Figure 4. It can be seen that the performance of six models still does not decline obviously with the increase of missing level, but different degrees of fluctuation appear, which is similar to that in MAR mechanism. At the same time, Figure 4 also shows that the performance fluctuation of ODCEM is minimal, which demonstrates that the ODCEM model proposed in this study has the best robustness for the MNAR type MVs in “German” dataset.

After analyzing the credit scoring performance of six models in “German” dataset with three missing mechanisms comprehensively, the following conclusions can be obtained:

(1) the overall performance of ODCEM model is always the best under three missing mechanisms;

(2) the MVs of different missing mechanisms can bring various degree effects on the performance of six models, and it is the greatest under MAR missing mechanism. This may be related to the production ways of MVs under three missing mechanisms.

4.5. Experimental Results Analysis in “China Churn” Dataset. Table 6 shows the customer churn prediction performance of six models in “China churn” dataset. It can be seen from the table that the total accuracy, AUC, type I accuracy, and type II accuracy of ODCEM are the best, which shows that ODCEM model can also achieve satisfactory performance in the real CVS dataset. It is notable that the performance of LMF is only comparable with that of RI-SVM, and their performance is the poorest in six models. “China churn” dataset contains a large number of MVs, and its average missing rate of instance $\overline{\alpha}$ is 8.06%. Combining with the analysis in Section 4.4, we can roughly conclude that the performance of ODCEM model proposed in this study is better than that of LMF model proposed by Mohammed et al. [34] when there are many MVs in the dataset.



(a) Missing level (%)

(b) Missing level (%)

(c) Missing level (%)

(d) Missing level (%)

Figure 4: The trend of classification performance in "German" dataset with MNAR type MVs.

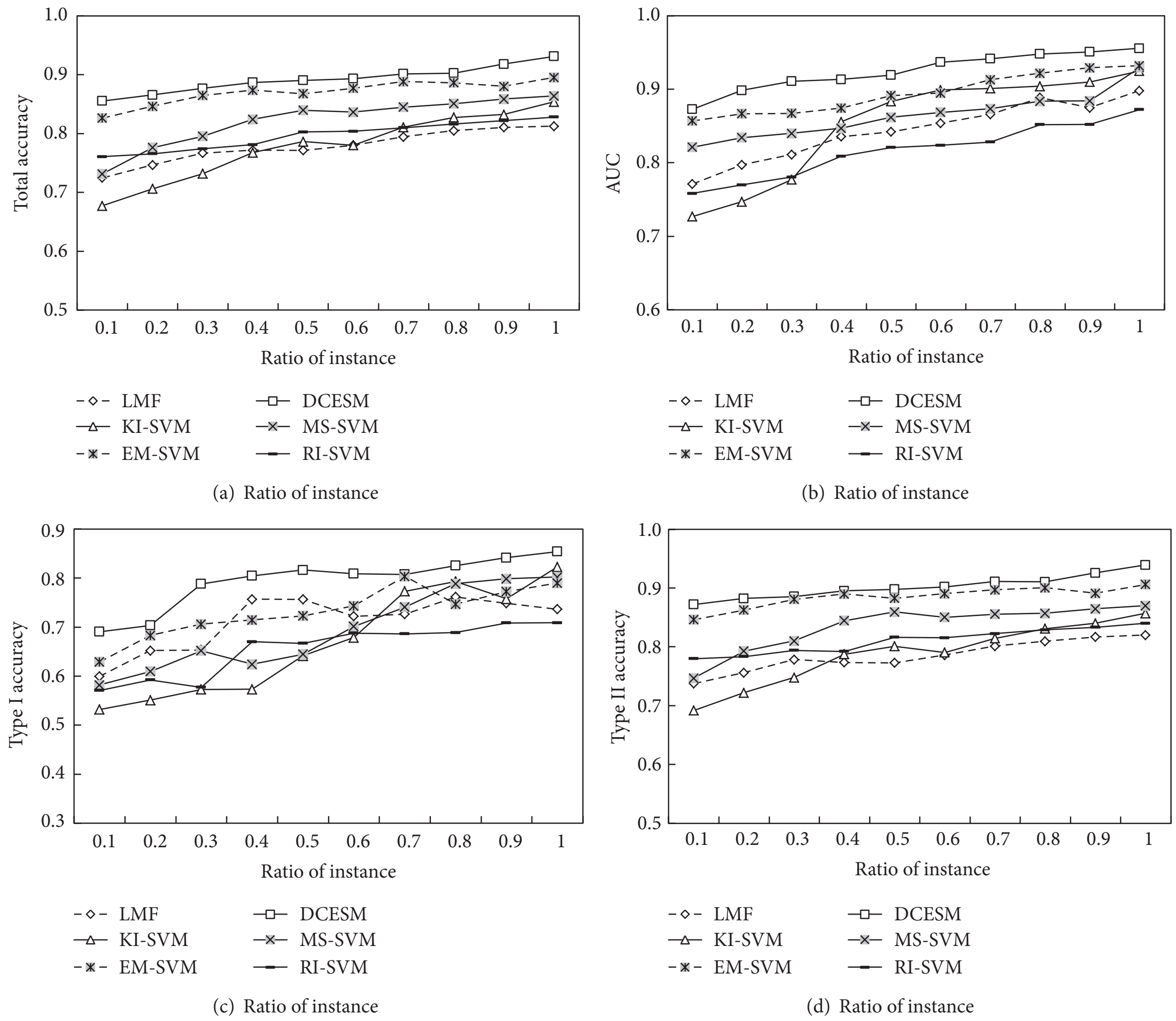
Further, the performance of six models in "China churn" dataset is shown in Figure 5. It can be seen that the performance of six models rises gradually when the ratio of instance for training model in the entire training set (abbreviated as ratio of instance) increases. Especially, the curve of ODCEM is always at the top even when the ratio is very small, such as 10%. It demonstrates that compared with the other five models, the churn prediction performance of ODCEM is the best when the models are trained by 10 training subsets with different sample size.

4.6. Further Discussions. In Sections 4.4 and 4.5, the experiments in "German" dataset and "China churn" dataset verify the effectiveness of the proposed method. Through accuracy and AUC, we can find which model is the best and which one is the poorest. However, the differences between the good models and bad ones are unclear [41]. Therefore, we conducted McNemar's test [42] to examine whether the proposed ODCEM model significantly outperforms the other five models referred to in this study. Taking the real customer churn prediction dataset "China churn" as an example, the results of McNemar's test are shown in Table 7. For space consideration, the results of McNemar's test for the "German" dataset with three missing mechanisms and five missing levels are omitted here. Actually, some similar conclusions can be obtained from "German" dataset by McNemar's test.

As shown in Table 7, it can be concluded as follows.

(1) The proposed one-step ensemble selection model ODCEM outperforms the "two-step" models, as well as the ensemble based model LMF at 1% statistical significance level.

(2) For LMF model, its performance is significantly poorer than that of EM-SVM at 1% statistical significance level and significantly poorer than that of MS-SVM and KI-SVM at 5% statistical significance level, while McNemar's test does not conclude that it performs poorer than the RI-SVM model.



(a) Ratio of instance

(b) Ratio of instance

(c) Ratio of instance

(d) Ratio of instance

Figure 5: Customer churn prediction performance of six models in “China churn” dataset. Note: x-axis shows the ratio of the instance for training model in the entire training set.

(3) For KI-SVM model, its performance is significantly better than that of RI-SVM at 10% statistical significance level and significantly poorer than that of EM-SVM at 5% statistical significance level. Finally, KI-SVM cannot outperform MS-SVM at 10% statistical significance level.

(4) For MS-SVM model, it outperforms RI-SVM at 10% statistical significance level. However, there is no significant difference between the performances of MS-SVM and EM-SVM at 10% statistical significance level.

Further, we also compare the computation complexity of six models and find that the complexity of MS-SVM is the lowest, followed by RI-SVM, KI-SVM, EM-SVM, LMF, and ODCEM. The time complexity of ODCEM and LMF is much the same, and it is slightly higher than that of EM-SVM.

Finally, with the analysis of Sections 4.4 and 4.5, we can draw the following conclusions.

(1) The CVS performance of ODCEM is the best in both UCI customer credit scoring dataset and real customer churn prediction dataset, which shows that ODCEM has good adaptability and can be used for a variety of CVS tasks.

(2) For the four “two-step” models, their performance is much different in the same dataset with three missing mechanisms and in different datasets (“German” and “China churn”), and the results are unstable. It demonstrates that the customer value classification modeling is sensitive to the preprocess methods of MVs in “two-step” CVS strategies and the CVS performance depends on the missing mechanism, which is similar to the conclusion of Crone et al. [24].

(3) For LMF model, when there are only a few MVs, such as in “German” dataset with the missing mechanism of MAR and MNAR, it achieves comparable performance with KI-SVM, MS-SVM, and EM-SVM,

Table 5: Comparison of performance in “German” dataset with MNAR type MVs.

Missing level	Evaluation criteria	LMF	ODCEM	KI-SVM	MS-SVM	EM-SVM	RI-SVM
$\theta = 5\%$	Total accuracy	0.7423 (3)	0.7467 (2)	0.7333 (5)	**0.7533** (1)	0.7400 (4)	0.7200 (6)
	AUC	0.8727 (4)	0.8740 (2)	0.8599 (6)	0.8731 (3)	**0.8923** (1)	0.8712 (5)
	Type I accuracy	0.7011 (2)	0.6821 (3)	0.6484 (6)	0.6564 (5)	**0.7077** (1)	0.6769 (4)
	Type II accuracy	0.7533 (5)	0.7744 (2)	0.7697 (3)	**0.7949** (1)	0.7538 (4)	0.7385 (6)
$\theta = 10\%$	Total accuracy	0.7300 (4)	**0.7567** (1)	0.7467 (2)	0.7333 (3)	0.7268 (5)	0.7000 (6)
	AUC	0.8314 (4)	**0.8577** (1)	0.8481 (2)	0.8434 (3)	0.8267 (5)	0.8200 (6)
	Type I accuracy	0.6411 (3)	**0.7011** (1)	0.6800 (2)	0.6333 (4)	0.5633 (6)	0.6000 (5)
	Type II accuracy	0.7510 (4)	**0.7805** (1)	0.7752 (3)	0.7762 (2)	0.7429 (5.5)	0.7429 (5.5)
$\theta = 20\%$	Total accuracy	0.7186 (3)	**0.7490** (1)	0.7069 (5)	0.7133 (4)	0.7269 (2)	0.7067 (6)
	AUC	0.8208 (3)	**0.8588** (1)	0.8175 (4)	0.8007 (6)	0.8393 (2)	0.8092 (5)
	Type I accuracy	0.6176 (3)	0.6524 (2)	0.5803 (5)	0.5837 (4)	**0.6574** (1)	0.5511 (6)
	Type II accuracy	0.7589 (5)	**0.7904** (1)	0.7611 (4)	0.7689 (3)	0.7567 (6)	0.7733 (2)
$\theta = 30\%$	Total accuracy	0.7433 (3)	**0.7600** (1)	0.7567 (2)	0.7367 (4)	0.7300 (5)	0.7200 (6)
	AUC	0.8498 (4)	**0.8763** (1)	0.8526 (3)	0.8490 (5)	0.8533 (2)	0.8244 (6)
	Type I accuracy	0.6222 (4)	**0.6945** (1)	0.6778 (2)	0.6289 (3)	0.5667 (5)	0.5556 (6)
	Type II accuracy	0.7952 (2)	0.7881 (5)	0.7905 (3.5)	0.7829 (6)	**0.8000** (1)	0.7905 (3.5)
$\theta = 40\%$	Total accuracy	0.7033 (5)	**0.7567** (1)	0.7367 (2)	0.7300 (3)	0.7167 (4)	0.7000 (6)
	AUC	0.8406 (4)	**0.8744** (1)	0.8486 (2)	0.8453 (3)	0.8382 (5)	0.8098 (6)
	Type I accuracy	0.6314 (4)	**0.6859** (1)	0.6585 (3)	0.6612 (2)	0.6266 (5)	0.6203 (6)
	Type II accuracy	0.7342 (5.5)	**0.7870** (1)	0.7702 (2)	0.7595 (3)	0.7553 (4)	0.7342 (5.5)
Average rank		3.73	1.50	3.33	3.40	3.68	5.38

Note: the bold-face in Table 5 shows the maximum of each row. The numbers in parentheses are the ranks of the six models with the corresponding evaluation criterion in each row.

Table 6: Comparison of churn prediction performance in “China churn” dataset.

Model	Total accuracy	AUC	Type I accuracy	Type II accuracy
LMF	0.8124	0.8480	0.7364	0.8201
ODCEM	**0.9313**	**0.9057**	**0.8539**	**0.9391**
KI-SVM	0.8538	0.8747	0.8229	0.8570
MS-SVM	0.8638	0.8793	0.8022	0.8701
EM-SVM	0.8954	0.8820	0.7899	0.9061
RI-SVM	0.8280	0.8224	0.7088	0.8401

Note: the bold-face in Table 6 shows the maximum of each column.

especially under the mechanism of MAR; its performance is only poorer than that of ODCEM. However, when there are many MVs, for example, in “German” dataset with the missing mechanism MCAR and “China churn,” its performance is poor. It shows that LMF model is not suitable for the CVS issues with lots of MVs, which is basically consistent with the experimental results of Mohammed et al. [34].

5. Conclusions and Future Work

This study mainly focuses on the CVS issues with MVs and proposes one-step dynamic classifier ensemble model (ODCEM) for MVs to make up for the disadvantage of the existing “two-step” models. On the one hand, ODCEM model integrates the preprocess of MVs and the classification modeling into one step; on the other hand, it utilizes multiple classifier ensemble technology in constructing the classification models. It can fully utilize the information of nonmissing values in dataset without imputation, thus reducing the dependence on the data missing mechanism assumptions and the data distribution. The empirical results in “German” dataset of UCI and the real customer churn prediction dataset “China churn” show that the CVS performance of ODCEM is better than that of four commonly used “two-step” models and the existing ensemble based model LMF.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Table 7: McNemar's test for pairwise comparison of performance in "China churn" dataset.

Model	LMF	KI-SVM	MS-SVM	EM-SVM	RI-SVM
ODCEM	15.254 (0.0000)	13.136 (0.0003)	12.971 (0.0003)	11.529 (0.0007)	14.469 (0.0001)
LMF		3.0480 (0.0809)	4.0364 (0.0367)	11.256 (0.0008)	0.0280 (0.8676)
KI-SVM			0.1880 (0.6650)	4.6880 (0.0304)	3.1840 (0.0744)
MS-SVM				2.7130 (0.1102)	3.5000 (0.0614)
EM-SVM					11.726 (0.0006)

Note: the results listed in Table 7 are the Chi squared values and *P* values are in brackets.

Acknowledgments

This research is partly supported by the Natural Science Foundation of China under Grant nos. 71101100, 71273036, and 71211130018, New Teachers' Fund for Doctor Stations, MOE (Ministry of Education), under Grant no. 20110181120047, Excellent Youth fund of Sichuan University under Grant no. 2013SCU04A08, China Postdoctoral Science Foundation under Grant nos. 2011M500418 and 2012T50148, Frontier and Cross-Innovation Foundation of Sichuan University under Grant no. skqy201352, Soft Science Foundation of Sichuan Province under Grant no. 2013ZR0016, Youth Project of Humanities and Social Sciences, MOE, under Grant no. 13YJC630249, and Scientific Research Starting Foundation for Young Teachers of Sichuan University under Grant no. 2012SCU11013.

References

[1] L. Ryals, "Making customer relationship management work: the measurement and profitable management of customer relationships," *Journal of Marketing*, vol. 69, no. 4, pp. 252–261, 2005.

[2] P. C. Verhoef and B. Donkers, "Predicting customer potential value an application in the insurance industry," *Decision Support Systems*, vol. 32, no. 1, pp. 189–199, 2001.

[3] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4176–4184, 2009.

[4] J.-B. E. M. Steenkamp and F. Ter Hofstede, "International market segmentation: issues and perspective," *International Journal of Research in Marketing*, vol. 19, no. 3, pp. 185–213, 2002.

[5] R. Venkatesan and V. Kumar, "A customer lifetime value framework for customer selection and resource allocation strategy," *Journal of Marketing*, vol. 68, no. 4, pp. 106–125, 2004.

[6] Y. Z. Liu and W. U. Hao, "A summarization of customer segmentation methods," *Journal of Industrial Engineering and Engineering Management*, vol. 20, no. 1, pp. 53–57, 2006.

[7] J. Lu, "Modeling customer lifetime value using survival analysis-an application in the telecommunications industry," Data Mining Techniques SUGI 28, 2008.

[8] S. A. Neslin, S. Gupta, W. Kamakura, L. U. Junxiang, and C. H. Mason, "Defection detection: measuring and understanding the predictive accuracy of customer churn models," *Journal of Marketing Research*, vol. 43, no. 2, pp. 204–211, 2006.

[9] A. K. Reichert, C. C. Cho, and G. M. Wagner, "An examination of the conceptual issues involved in developing credit-scoring models," *Journal of Business and Economic Statistics*, vol. 1, no. 2, pp. 101–114, 1983.

[10] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques," *Expert Systems with Applications*, vol. 34, no. 1, pp. 313–327, 2008.

[11] X. Hu, "A data mining approach for retailing bank customer attrition analysis," *Applied Intelligence*, vol. 22, no. 1, pp. 47–60, 2005.

[12] C.-F. Tsai and Y.-H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12547–12553, 2009.

[13] Y. Kim, "Toward a successful CRM: variable selection, sampling, and ensemble," *Decision Support Systems*, vol. 41, no. 2, pp. 542–553, 2006.

[14] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Applied Intelligence*, vol. 11, no. 3, pp. 259–275, 1999.

[15] C. K. Yim, D. K. Tse, and K. W. Chan, "Strengthening customer loyalty through intimacy and passion: roles of customer-firm affection and customer-staff relationships in services," *Journal of Marketing Research*, vol. 45, no. 6, pp. 741–756, 2008.

[16] A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 37, no. 5, pp. 692–709, 2007.

[17] S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: a case study," *Expert Systems with Applications*, vol. 31, no. 1, pp. 101–107, 2006.

[18] H. S. Subramania and V. R. Khare, "Pattern classification driven enhancements for human-in-the-loop decision support systems," *Decision Support Systems*, vol. 50, no. 2, pp. 460–468, 2011.

[19] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2002.

[20] S. Lessmann and S. Voß, "A reference model for customer-centric data mining with support vector machines," *European Journal of Operational Research*, vol. 199, no. 2, pp. 520–530, 2009.

[21] G. Paleologo, A. Elisseeff, and G. Antonini, "Subagging for credit scoring models," *European Journal of Operational Research*, vol. 201, no. 2, pp. 490–499, 2010.

[22] M. Li and L. Wang, "Feature fatigue analysis in product development using Bayesian networks," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10631–10637, 2011.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.

[24] S. F. Crone, S. Lessmann, and R. Stahlbock, "The impact of preprocessing on data mining: an evaluation of classifier

sensitivity in direct marketing," *European Journal of Operational Research*, vol. 173, no. 3, pp. 781–800, 2006.

[25] J. Van Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data and Knowledge Engineering*, vol. 68, no. 12, pp. 1513–1542, 2009.

[26] M. Ghannad-Rezaie, H. Soltanian-Zadeh, H. Ying, and M. Dong, "Selection-fusion approach for classification of datasets with missing values," *Pattern Recognition*, vol. 43, no. 6, pp. 2340–2350, 2010.

[27] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003.

[28] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[29] S. Wang, "Classification with incomplete survey data: a hopfield neural network approach," *Computers and Operations Research*, vol. 32, no. 10, pp. 2583–2594, 2005.

[30] J. W. Grzymala-Busse and L. K. Goodwin, "Handling missing attribute values in preterm birth data sets," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, D. Slezak, J. Yao, J. F. Peters, W. Ziarko, and X. Hu, Eds., vol. 3642 of *Lecture Notes in Computer Science*, pp. 342–351, 2005.

[31] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, NY, USA, 1987.

[32] D. B. Rubin, "Formalizing subjective notions about the effect of nonrespondents in sample surveys," *Journal of the American Statistical Association*, vol. 72, no. 359, pp. 538–543, 1977.

[33] S. Krause and R. Polikar, "An ensemble of classifiers approach for the missing feature problem," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 553–556, Portland, Ore, USA, July 2003.

[34] H. S. Mohammed, N. Stepenosky, and R. Polikar, "An ensemble technique to handle missing data from sensors," in *Proceedings of the IEEE Sensors Applications Symposium*, pp. 101–105, Houston, Tex, USA, February 2006.

[35] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[36] J. Xiao, C. He, X. Jiang, and D. Liu, "A dynamic classifier ensemble selection approach for noise data," *Information Sciences*, vol. 180, no. 18, pp. 3402–3421, 2010.

[37] C. Merz and P. Murphy, "UCI repository of machine learning databases," 1995, http://www.ics.uci.edu/~mlearn/MLRepository.html.

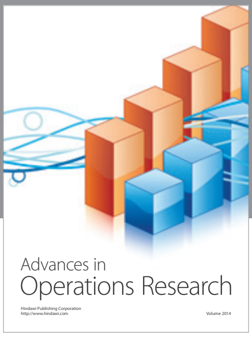
[38] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1027, 1993.

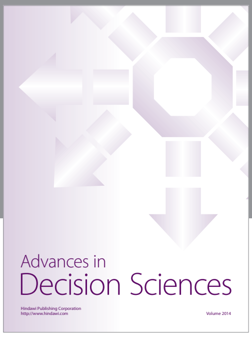
[39] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.

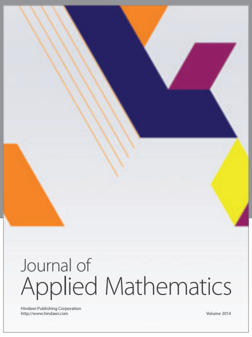
[40] T. Fawcett, *ROC Graphs: Notes and Practical Considerations for Researchers*, HPL-2004-03, Intelligent Enterprise Technologies Laboratory, Palo Alto, Calif, USA, 2004.

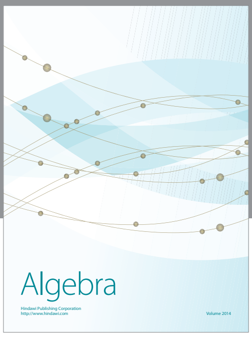
[41] L. Yu, S. Wang, and K. K. Lai, "An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: the case of credit scoring," *European Journal of Operational Research*, vol. 195, no. 3, pp. 942–959, 2009.

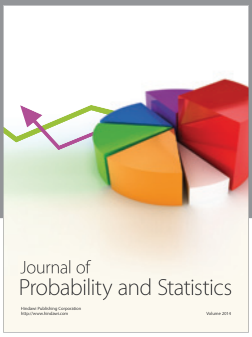
[42] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.


Advances in
Operations Research

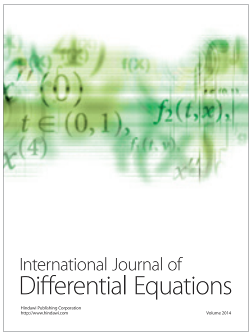

Advances in
Decision Sciences


Journal of
Applied Mathematics

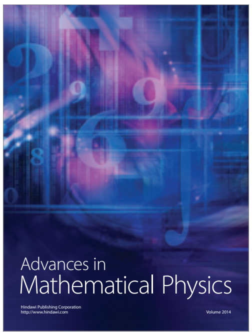

Algebra

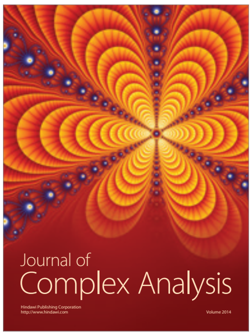

Journal of
Probability and Statistics

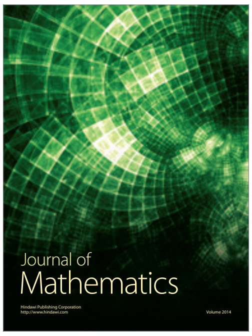
The Scientific
World Journal

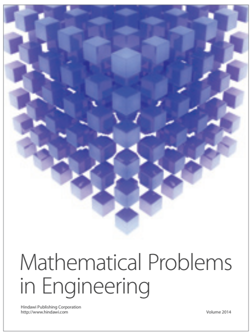

International Journal of
Differential Equations

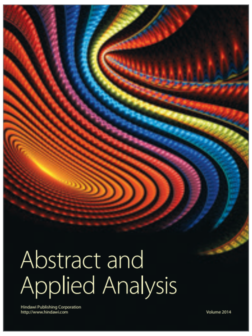
International Journal of
Combinatorics

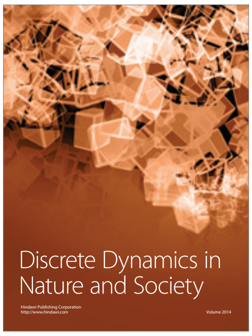

Advances in
Mathematical Physics

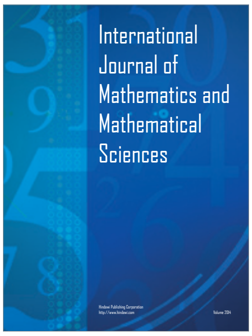

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

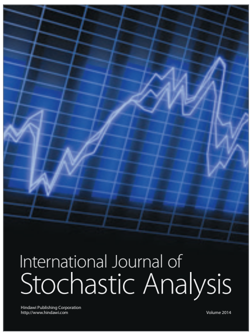

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization