Applied Soft Computing 10 (2010) 374–380

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Credit rating by hybrid machine learning techniques

Chih-Fong Tsai [a,*], Ming-Lun Chen [b]

[a] Department of Information Management, National Central University, 300 Jhongda Rd., Jhongli 32001, Taiwan
[b] Taichung Commercial Bank, Taiwan

ARTICLE INFO

Article history:
Received 30 May 2008
Received in revised form 15 April 2009
Accepted 2 August 2009
Available online 8 August 2009

Keywords:
Credit rating
Consumer loans
Machine learning
Hybrid models
Maximum profits

ABSTRACT

It is very important for financial institutions to develop credit rating systems to help them to decide whether to grant credit to consumers before issuing loans. In literature, statistical and machine learning techniques for credit rating have been extensively studied. Recent studies focusing on hybrid models by combining different machine learning techniques have shown promising results. However, there are various types of combination methods to develop hybrid models. It is unknown that which hybrid machine learning model can perform the best in credit rating. In this paper, four different types of hybrid models are compared by 'Classification + Classification', 'Classification + Clustering', 'Clustering + Classification', and 'Clustering + Clustering' techniques, respectively. A real world dataset from a bank in Taiwan is considered for the experiment. The experimental results show that the 'Classification + Classification' hybrid model based on the combination of logistic regression and neural networks can provide the highest prediction accuracy and maximize the profit.

1. Introduction

As the financial market competes sharply nowadays, the traditional banking profit is largely reduced. This causes banks to focus on consumer banking in order to make higher interest profits, i.e. consumer loans. However, the quality of issuing consumer loans is bank dependent and the audit process is forced to be as simple as possible. Consequently, the potential risk gradually arises.

With the rapid growth in credit industry and the management of large loan portfolios, credit rating (or credit scoring) models have been extensively used for the credit admission evaluation. The credit rating models are developed to classify loan customers as either a good credit group (accepted) or a bad credit group (rejected) with their related characteristics such as age, income and martial status or based on the data of the previous accepted and rejected applicants [1]. The benefits of considering credit scoring include reducing the cost of credit analysis, enabling faster decisions, insuring credit collections, and diminishing possible risks [24]. Even if a slight improvement in credit scoring accuracy might reduce large credit risks and translate into significant future savings.

The traditional approach to predict the consumers' credit risk is based on some statistical methods, such as logistic regression. However, related studies have shown that machine learning techniques or data mining techniques, such as neural networks, decision trees, etc., are superior to traditional (statistical) methods [2,9,22]. That is, using machine learning techniques can provide higher predication accuracy.

In machine learning, the hybridization approach has been an active research area to improve the classification/prediction performance over single learning approaches [8,11,13,17,19]. In general, it is based on combining two different machine learning techniques. For example, a hybrid classification model can be composed of one unsupervised learner (or cluster) to pre-process the training data and one supervised learner (or classifier) to learn the clustering result or vice versa [18].

Therefore, to develop a hybrid learning credit model, there are four different ways to combine the two machine learning techniques. They are: (1) combining two classification techniques, (2) combining two clustering techniques, (3) one clustering technique combined with one classification technique, and (4) one classification technique combined with one clustering technique.

In literature, related work developing credit rating models based on hybrid machine learning techniques only compare with some chosen single learning based models as the baselines (see Section 2.4). That is, none of the existing studies compares different hybrid models to identify which hybrid approach can perform the best for credit rating in terms of high prediction accuracy and low error rates.

Therefore, the aim of this paper is to examine the prediction performance of these four types of hybrid learning models for

* Corresponding author. Tel.: +886 3 422 7151; fax: +886 3 425 4604.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

1568-4946/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2009.08.003

credit rating in addition to the single classification and clustering techniques as the baseline models. Moreover, the profit made by these models is also compared based on a chosen bank as the case. Four well-known classification techniques, which are decision trees, Bayes classification, logistic regression, and neural networks, and two clustering techniques, which are *K*-means and expectation maximization are used to develop the hybrid models. Therefore, the contribution of this paper is to find out which combination method and techniques for the hybrid learning model can perform the best as well as provide maximum profits for credit rating.

The organization of this paper is as follows. Section 2 briefly describes related machine learning techniques used in this paper. In addition, related work is compared and their limitations are discussed. Section 3 presents the research methodology including the data used, the development of the credit rating models, evaluation strategies considered, etc. Section 4 shows the experimental results and the conclusion is provided in Section 5.

2. Machine learning techniques

2.1. Classification techniques

Classification (or supervised learning) techniques are based on learning by examples that map input vectors into one of several desired output classes. That is, a pattern classifier can be created through the training or learning process. The learning process of creating a classifier is to calculate the approximate distance between input–output examples and make correct output labels of the training set. This process is called the model generation phase. When the model is generated, it can classify an unknown instance into one of the learned classes in the training set [20].

2.1.1. Decision trees

A decision tree is a classification approach which is based on the tree structure to analyze data. One major advantage is that some decision rules can be produced that are easy to understand by humans. A decision tree classifies an instance by sorting it through the tree to the appropriate leaf node, i.e. each leaf node represents a classification. Each node represents some attribute of the instance, and each branch corresponds to one of the possible values for this attribute. C4.5 is the mostly used decision tree approach, and it is a later version of the ID3 algorithm [23].

2.1.2. Artificial neural networks

An artificial neural network, also called neural network, is composed of a group of neural nodes that link with the weighted nodes. Every node can simulate a neuron of creatures, and the connection among these nodes is equal to the synaptic that connects among the neurons. The most common type of neural networks consists of three layers of units: input layers, hidden layers, and output layers. It is called multilayer perceptron (MLP). A layer of "input" units is connected to a layer of "hidden" units, which is connected to a layer of "output" units [6].

2.1.3. Naïve Bayes classification

Naïve Bayesian classification [4] is based on Bayes theorem, which uses all kinds of beforehand probabilities and probabilities that are observed in the population to predict afterward probabilities. It is an effective tool to predict the relation of class members in the unknown situation.

The naïve Bayes classifier requires all assumptions be explicitly built into models which are then used to derive 'optimal' decision/classification rules. It can be used to represent the dependence between random variables (features) and to give a concise and tractable specification of the joint probability distribution for a domain. It is constructed by using the training data to estimate the probability of each class given the feature vectors of a new instance.

2.1.4. Logistic regression

Logistic regression is a simply parametric statistical approach. It is similar to traditional regression analysis. Therefore, the use of logistic regression should also conform to some hypothesis of traditional regression analysis, such as to avoid the autocorrelation in residuals, to avoid multi-collinearity in independent variables, and the collected data must conform to a normal distribution [7].

Logistic regression uses a series of numerical computations to establish a model by known classification parameters to find out which parameters have the more discriminated ability for each group and the classification rules for each group. Logistic regression is the same as discriminant analysis, which is used to deal with the relationship between independent variables and dependent variables when the dependent variable is a list of categories to which objects can be classified. Therefore, the difference between logistic regression and discriminant analysis is that discriminant analysis must satisfy the assumption of normal distribution and the equal covariance matrixes to find out the optimal value. However, logistic regression does not need these assumptions, even if these assumptions are satisfied, logistic regression can still provide relatively high prediction accuracy.

2.2. Clustering techniques

Clustering (or unsupervised learning) techniques can be regarded as the process of grouping similar objects into a cluster. In particular, labeled examples are not available. The purpose of clustering techniques is to improve the similarity of the members in a group and make the data in each cluster have the highest similarity, but the highest dissimilarity between different clusters [20].

Clustering algorithms can be classified into two categories, which are hierarchical and partitional clustering algorithms [12]. Hierarchical clustering creates a hierarchy of clusters by using the agglomeration algorithm. Then, a distinct singleton cluster will be combined one by one until satisfying some rules. The result will produce a series of arborescence partitions. On the other hand, partitional clustering is much more popular, which have been extensively used in many business problems [21]. Two well-known partitional clustering algorithms are *K*-means and expectation maximization (EM) algorithms described below.

2.2.1. K-means

The *K*-means clustering algorithm is a simple and efficient clustering method. *K*-means clustering is performed based on the following steps [5]:

- Given a group of feature vectors (or data points) as the dataset to be clustered.
- Randomly select the amount of seed by k to be the cluster center.
- Assign the nearest data points to the clusters.
- Average the position of every data point in the clusters in order to find out the new cluster, and then every data point will be assigned to their nearest cluster center.
- Repeat the two previous steps until some convergence criterion is met or the assignment cannot be changed.

2.2.2. Expectation maximization

The EM algorithm performs as the following steps [3]:

- *Step 1:* Estimation. First, assume the average of cluster parameter μ^i; standard deviation σ^i. Then, we can figure out the probability of p^i that every points to cluster q^i, and $q^i = 1$ where i represents

the ith cluster. Consequently, the initial values of μ^i, σ^i, and p^i are used to calculate the cluster probability of every instance χ. Next, the probability and responsibility to repeatedly predict the value of μ^i, σ^i, and p^i are applied.

The responsibility is defined by the degree of every data point of a dataset associated with cluster q^i. The responsibility for any χ can be obtained by:

$$h_n^i = \frac{(\pi_i/\sigma_i^d)e^{(-1/2\sigma_i^2)||x_n-\mu_i||^2}}{\sum(\pi_j/\sigma_j^d)e^{(-1/2\sigma_j^2)||x_n-\mu_j||^2}}$$

The naïve Bayes is:

$$h_n^i \equiv P(q_n^i = 1|x_n) = \frac{P(q_n^i = 1)P(x_n|q_n^i = 1)}{\sum P(q_n^i = 1)P(x_n|q_n^i = 1)}$$

- *Step 2:* Maximization. According to the responsibility of every data point, the average μ and standard deviation σ of the clusters are recalculated. Then, the new average μ is the cluster center.

The average μ is based on the responsibility of every data point in their clusters by:

$$\mu_i = \frac{\sum_n h_n^i x_n}{\sum_n h_n^i}$$

As a result, the likelihood of cluster q^i can be obtained based on μ^i, σ^i, and p^i defined by:

$$P(X_n|q^i = 1) = \prod_{k=1}^{n} P(X_k q^i = 1)$$

2.3. Hybrid machine learning techniques

In general, the hybrid model is based on combining the clustering and classification techniques. In Lenard et al. [18], two combination methods are proposed as shown in Figs. 1 and 2, respectively.

For the first hybrid model, as clustering is the unsupervised learning technique, it cannot distinguish data accurately like supervised one. Therefore, a classifier can be trained at first, and its output is subsequently used as the input for the cluster to improve the clustering result [10].

On the other hand, the second hybrid model uses the clustering technique first in order to filter out unrepresentative data. That is, the data which cannot be clustered accurately can be regarded as noisy data. Then, the representative data, which are not filtered out by the clustering techniques, are used to train the classifier in order to improve the classification result [8].

Besides, there are two other strategies to combine two machine learning techniques. The first one is based on combining two different classification techniques, and the second for combining two different clustering techniques.

For combining two classification techniques, given an original dataset D which contains n training and testing examples, the aim of the first classifier is to 'pre-process' D for data reduction. That is, the correctly classified data D' by the first classifier are collected, where D' contains m examples ($m < n$ and $D' \in D$). Then, D' is used to train the second classifier. Given a new testing set S, the second classifier could provide better classification results than single classifiers trained by the original dataset D.

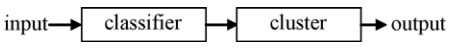

Fig. 1. A classifier combined with a cluster.

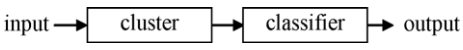

Fig. 2. A cluster combined with a classifier.

Similarly, for the combination of two clustering techniques, the fist cluster is also used for data reduction. The correctly clustered data D'' by the first cluster are used to train the second cluster, where D'' contains s examples ($s < n$ and $D'' \in D$). Finally, a new testing set S can be used to test the second cluster and other single clusters trained by the original dataset D for comparisons.

In short, the first component (no matter based on supervised or unsupervised learning technique) of the four hybrid approaches can simply perform the task of outlier detection [15]. That is, the D' and D'' datasets are much 'cleaner' than the original dataset D.

Note that some other sophisticated hybrid techniques, such as classifier ensembles [14] and stacking (or stacked generalization) [25] are not considered in this paper. This is because the computational efforts of constructing these models are relative higher than the above mentioned four hybrid approaches by combining two single techniques in a series connection. In particular, classifier ensembles need to combine a number of classifiers in parallel where the ensemble size is problem dependent and the outputs of these classifiers need to be further combined (or post-processed) by such as majority voting, or weighted voting, etc., for the final classification decision. On the other hand, the structure of stacking is much more complex than classifier ensembles, which is composed of n-level of classification where the first level classification is based on combining multiple classifiers parallelly and the second level classification is to collect the outputs of the first level classifiers as the training set to construct the second level classifier for the final classification. However, there is no a truly answer to the question about how many levels of the stacking scheme can perform the best [25].

2.4. Related work

Regarding Kumar and Ravi [16], which provide a detailed review of machine learning techniques for financial related problems, one important trend is to build a soft computing architecture (or hybrid intelligent systems). However, there are very few studies focusing on developing hybrid models for credit rating.

Table 1 compares related work of hybrid learning models in terms of their research problems, hybrid techniques used, and evaluation methods.

In related work, the developed hybrid models are generally compared with some chosen baseline models, which are based on single machine leaning techniques to make the final conclusion. Although they conclude that hybrid models outperform single classification models, it is unknown that what kind of the hybrid models can perform the best in the credit rating domain.

Moreover, much related work evaluates the models' performance by assessing their accuracy and error rates. However, no study examines the ability of making maximum profits of the credit rating models.

Therefore, this paper compares four different types of hybrid credit rating models in terms of their prediction accuracy, error rates, and the maximum profit they can make (c.f. Section 3.4).

3. Research methodology

3.1. The dataset

The dataset is collected from a business bank of Taiwan, which contains 12,929 cases of personnel consumer credit loan debit from 2004 to 2006. There are three different datasets used in this paper shown in Table 2. Table 3 shows their detailed information.

Table 1
Comparisons of related work.

Work	Problem domain	Hybrid techniques	Evaluation method
Hsieh [8]	Credit rating	Clustering + Classification	Accuracy and error rates
Huysmans et al. [10]	Credit rating	Classification + Clustering	Accuracy rates
Jain and Kumar [11]	Time series forecasting	Classification + Classification	Error rates
Kim and Shin [13]	Stock markets	Classification + Classification	Error rates
Lee et al. [17]	Credit rating	Classification + Classification	Accuracy and error rates
Malhotra and Malhotra [19]	Credit rating	Classification + Classification	Accuracy rates

Table 2
The three datasets.

Dataset ID	Description
Dataset 1	Contains the data whose proportion of the normal accounts and overdue accounts is 1:1. The data are selected randomly from the original dataset.
Dataset 2	Contains the original data without Dataset 1.
Dataset 3	Contains the data whose proportion of the normal accounts and overdue accounts is the same as the original dataset. The data are selected from Dataset 2.

Table 3
The dataset information.

Dataset	The ratio of normal and overdue accounts	No. of overdue accounts (cases)	No. of normal accounts (cases)
Original dataset	1:2.729	3467	9462
Dataset 1	1:1	1500	1500
Dataset 2	–	1967	7962
Dataset 3	1:2.729	1967	5368

In the original dataset, the numbers of normal accounts and overdue accounts are certainly different. In order to avoid ineffective model training, which may make the model inclines to the class containing the large quantity of data, we randomly select the data whose proportion of the normal accounts and overdue accounts is 1:1 from the original dataset as the dataset 1.

In particular, 10-fold cross-validation is used during the training and testing stages. That is, dataset 1 is divided into 10 un-duplicated subsets, in which any 9 of the 10 subsets are used for training and the remaining one for testing. As a result, each model will be trained and tested by 10 times.

To further validate the trained models, first of all we collect the original data without dataset 1, i.e. dataset 2, which are the unknown data for the trained models. Then, in order to make the validation set to be similar as the original dataset in terms of normal distribution, dataset 3 is produced from dataset 2 based on the same proportion of the normal accounts and overdue accounts as the original dataset.

3.2. Variables

The dataset contains a number of different input variables[1] shown in Table 4.

The most common variables are connubiality, query times from the joint credit information center, the quantity of cash cards, the quantity of credit card, liabilities, the job code, and age. The un-common variables, but are selected as significant variables in this dataset by the entropy algorithm are the case unit, the house ownership, the living class, the case source, delay of revolving credit, and the guarantor involved.

For the output variables, as credit rating can be regarded as a two-class classification problem, the model can classify or predict the new test case for either a good credit or bad credit.

[1] The representative variables are selected by the Microsoft SQL Server 2005 entropy algorithm.

Table 4
The variables in the dataset.

Case unit	Delay of revolving credit	The quantity of cash cards
House ownership	The guarantor involved (yes/no)	Query times form the joint credit information center
Living class	Connubiality	Liabilities
Case source	The quantity of credit card	Job code
Age		

Table 5
Parameter settings of the baseline models.

Model	Parameters
C4.5 decision tree	SCORE_METHOD = 4 SPLIT_METHOD = 3 MINIMUM_SUPPORT = 10 COMPLEXITY_PENALTY = default FORCED_REGRESSOR = default
Naïve Bayes	MINIMUM_DEPENDENCY_PROBABILITY = 0.5 MAXIMUM_STATES = 10
Logistic regression	HOLDOUT_PERCENTAGE = 30 HOLDOUT_SEED = 0 SAMPLE_SIZE = 1000 MAXIMUM_STATES = 100
Neural network	HIDDEN_NODE_RATIO = 4 HOLDOUT_PERCENTAGE = 30 HOLDOUT_SEED = 0 MAXIMUM_STATES = 100 SAMPLE_SIZE = 1000

3.3. Model development

There are six different types of credit rating models developed in this paper.[2] They are the four different types of hybrid models (c.f. Section 2.3) and two single baseline models based on the classification and clustering techniques individually.

3.3.1. Single models by classification techniques

The single baseline models using classification techniques are based on C4.5 decision trees, naïve Bayes, logistic regression, and neural networks, respectively. The parameter settings to construct the four baseline prediction models are shown in Table 5.

After running 10-fold cross-validation, the best baseline model can be obtained. Then, dataset 3 is used to evaluate the performance of these models.

3.3.2. Single models by clustering techniques

In this paper, there two clustering techniques used to develop the single baseline models. They are the *K*-means and expectation maximization (EM) algorithms. The number of clusters (i.e. the *k* value) is set from 2 to 50 to examine their clustering results based on 10-fold cross-validation. In particular, the percentages of the

[2] In this paper, the Microsoft SQL 2005 Data Mining software is used to develop the credit rating models.

normal and overdue accounts in every cluster are compared. Therefore, the cluster which contains a high percentage of normal accounts is used to recognize the group of normal accounts, and vice versa. Then, the best *K*-means and EM clusters are evaluated by dataset 3 to compare with other models.

3.3.3. Hybrid models by Classification + Clustering techniques

The best baseline classification model can be identified after performing 10-fold cross-validation using dataset 1, i.e. one of C4.5 decision trees, naïve Bayes, logistic regression, and neural networks. Then, the correctly predicted data from dataset 1 by the best baseline model are used as the new training data to train the best *K*-means and EM clusters, respectively. Therefore, the size of the correctly predicted data (i.e. the new training data) by the baseline model is smaller than dataset 1.

As a result, two hybrid models are developed. They are the best baseline model + *K*-means and the best baseline model + EM. Finally, dataset 3 is used to evaluate the prediction performance of the two Classification + Clustering hybrid models individually.

3.3.4. Hybrid Models by Classification + Classification techniques

Similar to the first stage of 'Classification + Clustering' techniques, the correctly predicted data from the best baseline classification model are used to train the four classification models individually, which are C4.5 decision trees, naïve Bayes, logistic regression, and neural networks.

Consequently, there are four hybrid models developed. They are (1) the best baseline classification model + C4.5 decision trees, (2) the best baseline classification model + naïve Bayes, (3) the best baseline classification model + logistic regression, and (4) the best baseline classification model + neural networks.

Again, dataset 3 is finally used to evaluate the prediction performance of the four Classification + Classification hybrid models individually.

3.3.5. Hybrid models by Clustering + Classification techniques

For *K*-means and EM, the cluster which provides higher accuracy by 10-fold cross-validation of dataset 1 is the best baseline clustering model. Next, the best clustering result (i.e. the accurately clustered data) from dataset 1 by the best baseline clustering model is used to train the four classification models individually. Therefore, the size of these data is also smaller than dataset 1.

As a result, four hybrid models are developed. They are: (1) the best baseline clustering model + C4.5 decision trees, (2) the best baseline clustering model + naïve Bayes, (3) the best baseline clustering model + logistic regression, and (4) the best baseline clustering model + neural networks.

Finally, dataset 3 is used to evaluate the prediction performance of the four Clustering + Classification hybrid models individually.

3.3.6. Hybrid models by Clustering + Clustering techniques

Similar to the type of 'Clustering + Classification' techniques, the best clustering result from the best baseline clustering model is used to further train *K*-means and EM, respectively. Therefore, in this type of hybrid models, the best baseline clustering model + *K*-means and the best baseline clustering model + EM are developed. Then, dataset 3 is also used to evaluate the prediction performance of the two Clustering + Classification hybrid models individually.

3.4. Evaluation methods

To evaluate the prediction performance of the developed credit rating models, prediction accuracy and Type I and II errors are considered. They can be measured by a confusion matrix shown in Table 6.

Table 6
Confusion matrix.

Actual	Predicted	
	Bad credit	Good credit
Bad credit	(a)	II (b)
Good credit	I (c)	(d)

Table 7
Prediction accuracy of baseline classification models.

Model	Testing dataset	Validation dataset	Ranking
DT	65.60%	72.97%	4
NB	71.20%	77.16%	3
LR	72.17%	79.42%	1
NN	71.60%	78.60%	2

Then, prediction accuracy can be obtained by:

$$\text{Prediction accuracy} = \frac{a+d}{a+b+c+d}$$

The Type I error shows the rate of prediction errors of a model, which incorrectly classifies the good credit group into the bad credit group. Opposed to the Type I error, the Type II error presents the rate of prediction errors of a model to incorrectly classify the bad credit group into the good credit group. Therefore, the Type II error is more critical which contain higher risks for financial institutions.

In addition, we will examine the breakeven rate which is based on the values in (a) and (b) of Table 6 to find out the optimal predication accuracy of the model which can provide maximum profits.[3] That is, the proportion of breakeven in this case bank is 5.351351:1, which means that the profit made from 5.351351 normal accounts can be counterbalanced by 1 overdue account.

4. Experimental results

First of all, average prediction accuracy of the six types of models (c.f. Section 3.3.) using the testing and validation datasets (i.e. dataset 1 and dataset 3) is present. Then, the six best models in each of the six different types of models are compared in order to find out which hybrid model performs the best in terms of prediction accuracy, Type I and II errors, and maximum profits made.

4.1. Prediction accuracy of the six types of credit rating models

4.1.1. Single baseline classification models

Table 7 shows average prediction accuracy of C4.5 decision trees (DT), naïve Bayes (NB), logistic regression (LR), and neural networks (NN), respectively. The result shows that the logistic regression model performs the best in this type of model.

4.1.2. Single baseline clustering models

Table 8 shows the prediction result of the best *K*-means and EM with their numbers of clusters, respectively. EM provides the highest accuracy when the number of clusters is 41 and *K*-means with 29 clusters performs the best. The comparative result shows that EM is the best single clustering model.

4.1.3. Classification + Clustering hybrid models

As shown above, the best single classification model is identified, which is logistic regression. It provides 72.17%

[3] We adjust the value of the 'predict probability' function provided by the Microsoft SQL 2005 Data Mining software to obtain maximum profits.

Table 8
Prediction accuracy of baseline clustering models.

Model	Testing dataset	Validation dataset	No. of clusters
EM	62.9%	68.45%	41
K-means	55.23%	56.9%	29

Table 9
Prediction performance of LR.

Actual	Predicted	
	Bad credit	Good credit
Bad credit	1169	504
Good credit	331	996

Table 10
Prediction performance of LR + EM and LR + *K*-means.

Model	Prediction accuracy	No. of clusters
LR + EM	76.07%	41
LR + *K*-means	70.37%	29

Table 11
Prediction performance of the four hybrid models.

Model	Prediction accuracy	Ranking
LR + DT	72.50%	4
LR + NB	78.68%	3
LR + LR	82.94%	2
LR + NN	83.44%	1

Table 12
Prediction performance of EM.

Actual	Predicted	
	Bad credit	Good credit
Bad credit	918	582
Good credit	531	969

prediction accuracy and has correctly predicted 2165 data from dataset 1, in which the normal and overdue accounts are 1169 and 996, respectively. Table 9 shows the prediction performance of logistic regression (LR) by a confusion matrix.

Then, the 2165 sample data are used to train the best *K*-means and EM shown in Table 8. Table 10 shows the prediction performance of these two hybrid models tested by dataset 3. In this type of hybrid model, the LR + EM hybrid model performs the best.

4.1.4. Classification + Classification hybrid models

Similar to above hybrid models, the 2165 data which are accurately predicted by LR are used to train DT, NB, LR, and NN, respectively. Table 11 shows the prediction performance of the four hybrid models using dataset 3. The LR + NN hybrid model provides the highest rate of prediction accuracy.

Table 13
Prediction performance of the four hybrid models.

Model	Prediction accuracy	Ranking
EM + DT	64.44%	4
EM + NB	72.28%	3
EM + LR	80.16%	1
EM + NN	74.63%	2

Table 14
Prediction performance of EM + EM and EM + *K*-means.

Model	Prediction accuracy	No. of clusters
EM + EM	66.65%	41
EM + *K*-means	63.83%	29

4.1.5. Clustering + Classification hybrid models

Regarding the result of single baseline clustering models, EM with 41 clusters performs the best. Table 12 shows the prediction performance of EM by a confusion matrix.

Therefore, 1187 data which are accurately predicted by EM are used to train DT, NB, LR, and NN, respectively. Table 13 shows the prediction performance of the four hybrid models using dataset 3. In this type of hybrid models, EM + LR provides the best prediction result.

4.1.6. Clustering + Clustering hybrid models

In this type of hybrid models, the 1187 data which are accurately predicted by EM are used to train *K*-means (with 29 clusters) and EM (with 41 clusters), respectively. Table 14 shows their prediction performances by dataset 3. The result indicates that the EM + EM hybrid model performs the best.

4.2. Comparisons of the six best credit rating models

Table 15 compares the six best models in the six types of credit rating modes in terms of their prediction accuracy, numbers of normal and overdue accounts, and profit units. The number in the bracket means the model's ranking.

Regarding this comparative result, the 'Classification + Classification' hybrid model performs the best, which provide 83.44% prediction accuracy. In particular, logistic regression used as the first component combined with neural networks as the second component, i.e. LR + NN, is superior to the other models. Moreover, it also can maximize the profit. On the other hand, LR + NN can accurately predict the most normal and overdue accounts, which means that it provides the lowest Type I and II errors. Therefore, this hybrid model can be regarded as the optimal credit rating system.

It is interesting that the 'Classification + Clustering' and 'Clustering + Clustering' hybrid models do not outperform single classification models. This implies that the clustering techniques cannot provide reasonable credit rating results. On the other hand, although the 'Clustering + Clustering' hybrid model performs the worst, it can provide higher profits than single classification and

Table 15
Performance of the six best models.

Technique	Model	Accuracy	Normal accounts	Overdue accounts	Profit units
Single Classification	LR	79.42% (3)	1904 (4)	19 (5)	336.93 (4)
Single Clustering	EM	68.79% (5)	729 (6)	15 (6)	121.03 (6)
Classification + Clustering	LR + EM	76.07% (4)	1842 (5)	20 (4)	324.21 (5)
Classification + Classification	LR + NN	83.44% (1)	3926 (1)	154 (1)	579.65 (1)
Clustering + Classification	EM + LR	80.16% (2)	3093 (2)	29 (3)	548.98 (2)
Clustering + Clustering	EM + EM	66.65% (6)	2529 (3)	97 (2)	375.59 (3)

clustering models and the 'Classification + Clustering' hybrid model, and it stands for the second place over the Type II error.

5. Conclusion

As there is some risks of issuing consumer loans by banks, developing credit rating systems is necessary. This paper focuses on comparing various hybrid machine learning models for the credit rating problem. In particular, there are four different types of hybrid models are developed. They are 'Classification + Clustering', 'Classification + Classification', 'Clustering + Classification', and 'Clustering + Clustering' hybrid models.

The experimental results show that the 'Classification + Classification' hybrid model provides the highest prediction accuracy and lowest error rates and can make the maximum profit. The optimal credit rating model is based on logistic regression as the first classifier combined with neural networks as the second classifier. On the other hand, clustering techniques (both single and hybrid models) used to produce the final decision cannot outperform the classification ones.

For practice, the finding of this paper allows us to understand which kind of hybrid approaches is able to produce higher rates of prediction accuracy and lower error rates in terms of credit rating. In addition, based on the optimal credit rating model identified in this paper (i.e. LR + NN) related financial institutions can make more correct decisions for issuing consumer loans with high confidence in the future.

It should be noted that although this paper considers several popular techniques to develop the four types of hybrid models, there are other algorithms available in literature, which can be applied, for example, self-organizing maps (SOM) as the clustering technique, and support vector machines and genetic algorithm for the classification techniques. However, from the practical standpoint, it is impossible to conduct a comprehensive study on all existing clustering and classification techniques. In particular, it is hard to define the most representative techniques in the credit rating domain and conduct a complete comparative study based on these techniques.

For future work, the ensemble approach [14] and stacked generalization [25] can be considered for further comparisons as they are based on combing multiple classification techniques in a parallel form. On the other hand, based on the characteristics of the consumer loans, they can be specifically divided into, such as car loans, house loans, personal small scale loans, etc., for individual studies. Finally, since this paper mainly focuses on the crediting rating problem, the experimental setup can be applied to other domain problems to find out which hybrid approach performs the best or if the experimental results are different from the findings of this paper.

Acknowledgement

This research is partially supported by National Science Council of Taiwan (NSC 96-2416-H-194-010-MY3).

References

[1] M.-C. Chen, S.-H. Huang, Credit scoring and rejected instances reassigning through evolutionary computation techniques, Expert Systems with Applications 24 (4) (2003) 433–441.
[2] J.N. Crook, D.B. Edelman, L.C. Thomas, Recent developments in consumer credit risk assessment, European Journal of Operational Research 183 (2007) 1447–1465.
[3] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B 39 (1) (1977) 1–38.
[4] D.G.T. Denison, C.C. Holmes, B.K. Mallick, A.F.M. Smith, Bayesian Methods for Nonlinear Classification and Regression, Wiley, 2002.
[5] J.A. Hartigan, M.A. Wong, A *K*-means clustering algorithm, Applied Statistics 28 (1) (1979) 100–108.
[6] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice Hall, New Jersey, 1999.
[7] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd ed., Wiley, New York, 2000.
[8] N.-C. Hsieh, Hybrid mining approach in the design of credit scoring models, Expert Systems with Applications 28 (2005) 655–665.
[9] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, S. Wu, Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision Support Systems 37 (2004) 543–558.
[10] J. Huysmans, B. Baessens, J. Vanthienen, T.V. Gestel, Failure prediction with self organizing maps, Expert Systems with Applications 30 (2006) 479–487.
[11] A. Jain, A.M. Kumar, Hybrid neural network models for hydrologic time series forecasting, Applied Soft Computing 7 (2) (2007) 585–592.
[12] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Computing Survey 31 (3) (1999) 264–323.
[13] H.-J. Kim, K.-S. Shin, A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets, Applied Soft Computing 7 (2) (2007) 569–576.
[14] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.
[15] E.M. Knorr, R.T. Ng, Algorithms for mining distance-based outliers in large databases, in: Proceedings of the 24th VLDB Conference, 1998, pp. 392–403.
[16] P.R. Kumar, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review, European Journal of Operational Research 180 (2007) 1–28.
[17] T.-S. Lee, C.-C. Chiu, C.-J. Lu, I.-F. Chen, Credit scoring using the hybrid neural discriminant technique, Expert Systems with Applications 23 (2002) 245–254.
[18] M.J. Lenard, G.R. Madey, P. Alam, The design and validation of a hybrid information system for the auditor's going concern decision, Journal of Management Information Systems 14 (4) (1998) 219–237.
[19] R. Malhotra, D.K. Malhotra, Differentiating between good credits and bad credits using neuro-fuzzy systems, European Journal of Operational Research 136 (2002) 190–211.
[20] T. Mitchell, Machine Learning, McGraw Hill, New York, 1997.
[21] S. Olafsson, X. Li, S. Wu, Operations research and data mining, European Journal of Operational Research 187 (2008) 1429–1448.
[22] C.-S. Ong, J.-J. Huang, G.-H. Tzeng, Building credit scoring models using genetic programming, Expert Systems with Applications 29 (2005) 41–47.
[23] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[24] D. West, Neural network credit scoring models, Computers and Operations Research 27 (11/12) (2000) 1131–1152.
[25] D.H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259.