Assessment of Customer Credit with Efficient Multivariate Classifier

HUNG-YI LIN[1] & KUANG-YU HUANG[2]
[1]Department of Distribution Management
National Taichung University of Science and Technology
129, Sanmin Rd., Sec. 3, Taichung
Taiwan (R.O.C.)
linhy@nutc.edu.tw

[2]Department of Information Management
Ling Tung University
1, Ling Tung Rd., Taichung
Taiwan (R.O.C.)
kyhuang@teamail.ltu.edu.tw

Abstract: - The assessment of customer credit is important for financial institutions. Many techniques are developed for customer credit classification. Traditional methods suffer from the problems of inaccurate prediction and/or inefficient data analysis. In this paper, we adopt the entropy-based evaluation method and PCA-based classifier to improve these two problems. A new feature evaluation criterion collaborated with a novel feature selection are proposed to identify the critical factors of classification. By means of multivariate analysis, not only a reduced subset of relevant features is achieved but also the efficient classifier is generated. This paper aims at assessing customer credit with effectiveness and efficiency. From the experimental results of German credit case, it shows that our propsed method can simultanously improve accuracies in the classes of high-credit and bad-cedit when competing with the traditional C4.5 scheme.

Key-Words: - Customer credit assessment, feature selection, multi-class classification, PCA, multivariate analysis, C4.5

1 Introduction

After the subprime mortgage crisis emerged from United States, worldwide enterprises and commercial communities have intensified their efforts to assess customer credits. This is because the inaccurate customer credit assessment may cause financial institutions run into financial difficulty and lead to heavy costs in afterwards management. Customer credit assessing is the important basis for developing the lending strategies and becomes one of the most challenging tasks and a research topic in CRM [21].

The ubiquitous information tools and technology allows us to collect and store a variety of questionnaire and personal data about credit and financial statuses with ease. The ripe development of multi-dimensional database management, artificial intelligence, expert system, and data mining techniques make easy implementation of various investigation, studies, or analyses on the stored data. Decision makers have gained more momentum to tackle difficult problems due to modern computer technologies and computational techniques.

Classification is a problem frequently encountered when a categorical dependent variable needs to be predicted according to a small subset of independent variables. Many classification problems including web page classification [22], web spam detection [3, 6], intrusion detection [7], mobile commerce behavior [19], fraud detection [5], bankruptcy prediction [24], medical diagnosis [9, 13], and crime activity analysis [11], have attracted many research attentions. Recent studies [12, 20, 28] have developed multivariate classification methods to improve accuracy or promote classification performance. We note that high performance is especially important since modern data amount is increasingly growing with a high speed.

Complexity, efficiency, and accuracy are three principles in appraising classifiers. The complexity of one classifier relies on the amount of involved variables and the processing of data analyses. Generally, classifier complexity is closely related to classification efficiency. In other words, one classifier with simple handle can complete the classification task with high efficiency. However, it is a general case that losing classification efficiency in return can gain classification accuracy and vice

versa. This trade-off is a well-known challenge in classification problems. There is no doubt that the approach to simultaneously improving efficiency and accuracy is highly expected.

On the one hand, in order to collect sufficient information, every instance usually includes a lot of categorical and/or numerical variables (i.e., features or factors). These variables generally reflect independent data elements. For clustering problem, all variables are considered independently. For classification problem, one or few variables are taken as dependent for prediction purpose. On the other hand, in order to induce data trend and then deduce future data, it is also required to have an enough amount of instances for analyses. In this scenario, overabundance becomes the bottleneck when filtering and condensing the great amount of data. Note that feature selection and typical instance recognition are two essential issues to remedy such problems. We only focus on the first issue in this paper.

Feature selection aims at exploring the effective variables in data mining process. Many statistical and artificial intelligence techniques are devoted to identify a set of discriminative features. These techniques include genetic algorithm [8, 27], support vector machine [4, 26], neural network and fuzzy [18, 23, 25], logistic regression [10, 15], and principal component analysis [16]. The objective of feature selection is threefold. First is to eliminate the redundant information so that the analytical time of mining process can be reduced. Second, the selection of a small subset of low correlated features will facilitate data mining process since it prevents similar factors from being repeatedly involved. Thirdly, for classification problems, the relevant features to the target feature are more effective than the irrelevant ones when connecting their relationships.

In order to explore the useful data attributes in assessing customer credit, an entropy-based criterion is proposed in evaluating all input variables in this paper. This criterion takes the data distribution and data variation of the target feature into consideration. This criterion is suitable to evaluate categorical and numerical input features. After a small subset of discriminative features is identified, a multivariate classifier is then generated according to the statistical tool of principal component analysis. Customers with different attribute values can be efficiently classified into several distinct credit levels.

This paper is organized as follows. Section 2 first proposes the new learning model and then addresses the novelties in this model. Section 3 explains the issues for multiple classes and data granularity. The detail procedures for generating the multivariate classifier are given in Section 4. Section 5 describes the real case of German credit and the experimental results are presented as well. The conclusion is provided in Section 6.

2 New Inductive Learning Model

Figure 1 shows a new inductive learning model constituted with three stages: 1) feature selection, 2) multivariate analysis, and 3) class matching. After a collection of training data is inputted to the first stage, data preprocessing such as standardizing, z-scoring, or normalizing is first applied for overcoming the problem attributed to different measure scales. Then, all features are evaluated by using the processed data in classifying the target classification variable. All evaluation values are sorted and the features with high ranks are selected for the subsequent handles. Principal component

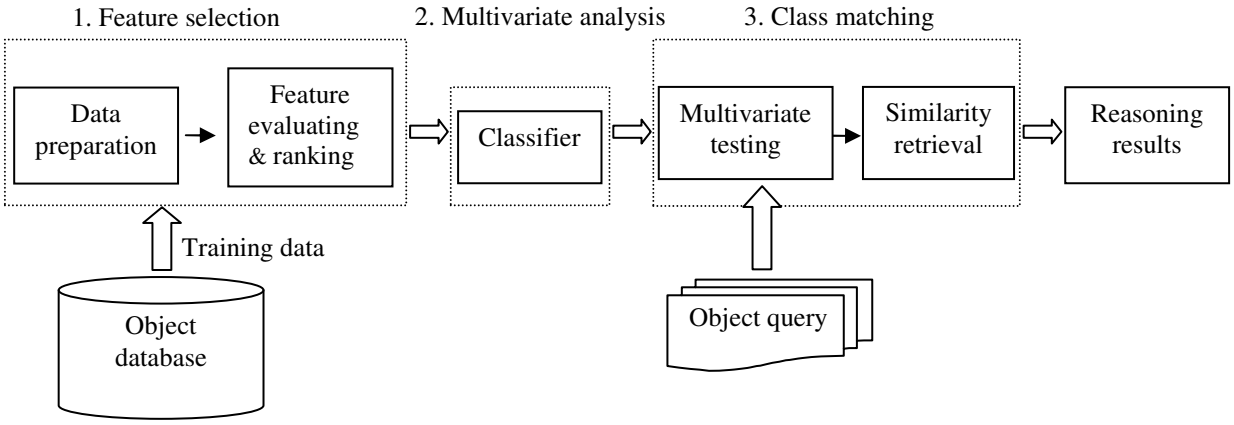


Fig. 1 Inductive learning model.

analysis is applied on the selected features in the second stage and a multivariate classifier is constructed for further handles. In the third stage, the resulting classifier is trained for class matching. Eventually, the well-trained classifier is built. New data object waiting for classification is predicted by the classifier. The detailed procedures for these stages will be described in Section3.

We emphasize the novelties of our new inductive model as follows. First of all, the design of Shannon's *Information Theory* is revised for the purpose of data diversity reduction. *Entropy* and *Information Gain* calculate the information amount in every subclass and then integrates them for the overall measure. Note that the common concept of these entropy-based designs concentrates on data distribution (i.e., the statistical dispersion of data). In case of multi-class classification problems, a higher variety of classes is employed in the target feature. It is problematic if data variation inside every subclass is ignored. We clearly explain such problem in the next section.

The second stage implements an inductive scheme which extracts the relevant information for the facilitation of efficient classification. To grasp the significant factors without the sacrifice of classification effect, the methods of *explained variance* and *communality* in principle component analysis (PCA) are carefully investigated. The goal of this stage is to train a new multivariate classifier.

For every queried object, the third stage takes the advantage of the resulting multivariate classifier for similarity retrieval. The final goals of this model are to complete efficient reasoning task and then quick respond dataset users with the high quality of query answers.

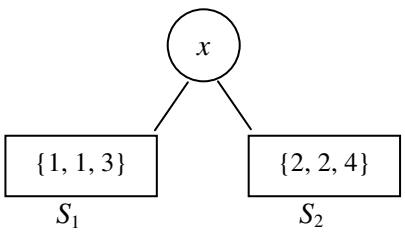
3 Issues for Multiple Classes and Data Granularity

The main concept of Shannon's *information theory* quantifies the uncertainty involved in predicting the value of a random variable. Information theory is developed from probability theory and statistics. The most important quantities of information are entropy, which measures the amount of uncertainty associated with a discrete random variable. *Information Gain* is an entropy-based way of comparing two distributions in a manner that assumes $q(X)$ to be the distribution underlying some data and $p(X)$ to be the correct distribution. It is thus defined as $IG(p(X)|q(X)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$. *Information Gain Ratio* adds the factor of population information amount. These entropy-based criteria all focus on evaluating the random variables with two classes (i.e., yes vs. no, or high vs. low, or good vs. bad). Although the analyses of data distribution (i.e., the statistical dispersion of data) can help distinguish the quality of classification effect, classification problems involving multiple classes tend to have large data variation inside every subset. The following example is used to illustrate this concern.

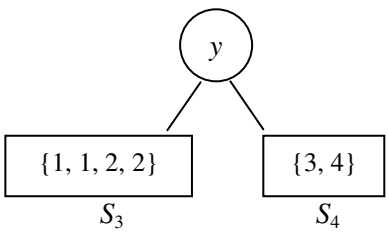
As shown in Fig. 2(a), a small dataset has two independent variables (x, y) and four classes in its target classification variable c. After classifying all instances using x and y, the resulting subsets are depicted in Figs. 2(b) and 2(c), respectively. Intuitively, we vote x as the better feature than y since the entropy gained from x (i.e., 0.92) is better than that from y (i.e., 1). However, the average data variance of S_1 and S_2 (i.e., 0.89) is greater than that of S_3 and S_4 (i.e., 0.25). Or, the data distribution in S_1 has the probabilities of 0.67 (c = 1) and 0.33 (c = 3), which have higher consistency than that in S_3, the probabilities of 0.5 (c = 1) and 0.5 (c = 2). The same condition happens to S_2 and S_4. However, data

x	y	c
0	0	1
0	0	1
1	0	2
1	0	2
0	1	3
1	1	4

(a)



(b)



(c)

Fig. 2 An example with six instances.

variation in S_1 (and S_2) is greater than that in S_3 (and S_4). As is well known, feature evaluation methods using the criterion of data distribution concern only the *impurity* inside each subset. Multiple classes make *proximity* another concern, which is ignored by many conventional evaluation criteria. Criteria such as entropy-based methods and Gini index are insufficient to evaluate the features in multi-class problems. Our new evaluation criteria will give consideration to both impurity and proximity. The detailed design is given in Section 4.

Data granularity refers to the fineness with which data fields can be sub-divided. For example, census data can be recorded, with low granularity or coarse-grained description, as a single statement: [one has 6 children at the age of 45 and his education level and standard of living are high], or with high granularity or fine-grained description, as multiple statements:

1. [Age = 45]
2. [Number of children = 6]
3. [Education level = high]
4. [Standard of living = high]

Conventional decision trees adopt the concept of fine granularity in every single decision node, which classifies instances level by level. The initial classification triggered by the root node determines the subsequent handling. An adequate high-level decision can simplify the classification tasks assigned at low levels. In the case of multi-class classification, every single variable pays close attention to its own data dimension. It is difficult to vote the most adequate variable and locate it in the root node for generating the optimal classification effect for all subsequent handling. As a result, different types of decision trees can build different classification models and achieve different effectiveness. In fact, the exploration of an optimal decision tree for multi-class problem is a NP-complete problem.

Now that a thoughtful classification is usually accomplished by the cooperation of a collection of relevant variables, data classification according to coarse granularity does not have a bias in favour of one single data dimension but simultaneously takes multiple data dimensions into account. That is, instead of applying univariate classification, the learning model with multivariate classifier can take more aspects into account and in turn achieve better classification quality.

4 Feature Selection Scheme

To enhance the classification performance, PCA is integrated into our proposed learning model. PCA is also referred to as Karhunen-Loève transform [14]. This transform is able to reproduce the total system variability and achieves high reduction in dimensionality with usually lower noise than the original patterns. The disadvantage of PCA is that the principle components are not easy to interpret. However, the gain of analytical performance is derived from this disadvantage. The mathematical background of PCA is explained as follows.

Suppose there are p input variables in a dataset. And, this dataset consists n measurements on these p variables. In general, often much of the total system variability can be accounted for by a small number k of the principal components. Namely, there is as much information in the k components as there is in the original p variables. The k principal components can then replace the initial p variables and the original data set.

Principal components are particular linear combinations of a set of random variables (respectively denoted as x_1, x_2, …, and x_p). Principal components depend solely on the covariance matrix Σ of these random variables. Their developments do not require a multivariate normal assumption. Let the eigenvalues of the covariance matrix Σ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. The i-th component is the linear combination which is given by $y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p$, where $[a_{i1}, a_{i2}, \cdots, a_{ip}]$ is the eigenvector corresponding to λ_i, $i = 1,2,\cdots, p$. Consequently, the proportion of total variance (i.e., explained variance) due to the i-th principal component is $\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$. The correlation coefficient between the component y_i and the variable x_k is $\rho_{i,k} = \frac{a_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_k^2}}$, where $i,k = 1,2,\cdots p$. If most (for instance, more than 70%) of the total population variance can be featured to the first or two components, then these components can replace the original p variables without much loss of information. In this paper, only the first principal component with maximum variance is considered as an effective use to replace the original p features.

Data distribution and data variation inside every class are measured to evaluate all input features. Entropy measure and statistical variance collaborate to identify classification effect. Based this design, a modest number m of input features are sifted from

all. The detailed steps in determining the value of *m* are shown in Fig. 3.

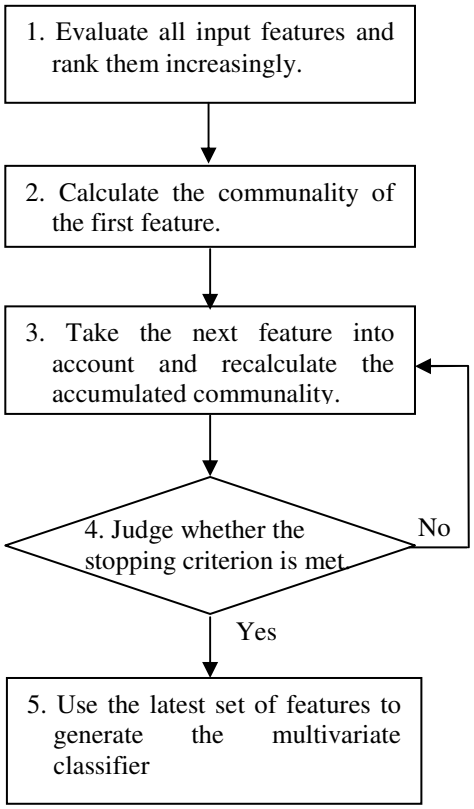


Fig. 3 Feature selection procedure.

Step 1. Evaluate all input features and rank them increasingly.

As mentioned above, not all input features are useful and effective for the facilitation of classification. Many of them are clumsy in generalizing the categorization from a collection of data. A revised entropy-based evaluation method called intra-diversity is proposed as follows.

Suppose a feature x with n possible values (numeric, nominal, or cardinal) is employed to classify a training set T. Then, T is subdivided into n subclasses, respectively denoted as T_1, T_2, …, T_n. The intra-diversity $ID(x,T)$ is formulated as

$$ID(x,T)=\sum_{i=1}^{n}\frac{Ent(T_i)\times Var(T_i)}{|T_i|} \quad (1)$$

, where Shannon entropy ($Ent(T_i)$) considers data distribution and variance ($Var(T_i)$) takes data variation into account. They collaborate to evaluate the data aggregation acquired in every subclass. In order to reduce the effect of the subclass size, the product is weighted by $|T_i|^{-1}$. The smaller estimation of $ID(x,T)$ indicates the less data diversity and in turn the more data aggregation inside every subclass.

Suppose there are p input features (x_1, x_2, …, x_p) in the database. After the *ID* of all input features are estimated, they are resorted into a new list based on an inceasing order ($x^{(1)}$, $x^{(2)}$, …, $x^{(p)}$). The ones ranked on the front side of the list are more effective than those ranked on the rear side. Hence, a modest number m of features are taken from the front side of the list as the representatives for further processing. We suggest m to be ($\lfloor\sqrt{p}\rfloor+1$) in this paper.

Step 2. Calculate the communality of the first feature.

After applying PCA on the m features extracted from Step 1, the correlation coefficient between componet y_1 and variable $x^{(k)}$ is $\rho_{y_1,x^{(k)}}=\frac{a_{1k}\sqrt{\lambda_1}}{\sqrt{\sigma_k^2}}$, where $k\in[1,2,\cdots,m]$. The communality of $x^{(1)}$ for component y_1 stands for the proportion of variance of $x^{(1)}$ that is due to the component y_1. It is measured as $\rho^2_{y_1,x^{(1)}}=\frac{a_{11}^2\lambda_1}{\sigma_k^2}$.

Step 3. Take the next feature into account and recalculate the accumulated communality.

The communality accumulated from $x^{(1)}$ and $x^{(2)}$ for component y_1 is evaluated as $A_2=\rho^2_{y_1,x^{(1)}}+\rho^2_{y_1,x^{(2)}}$. Generally, the communality accumulated from $x^{(1)}$ to $x^{(k)}$ for the component y_1 is formulated as $A_k=\sum_{j=1}^{k}\rho^2_{y_1,x^{(j)}}$, $k\in[2,3,\cdots,m]$.

Step 4. Judge whether the stopping criterion is met.

In case of k features are considered, PCA is applied again on these k features and then the proportion of total variance accounted for by the first component is evaluated as $B_k=\frac{\Lambda_1}{\sum_{i=1}^{k}\Lambda_i}$, where Λ_i is the i-th eigenvalue of the covariance matrix $\sum$ formed by these k features. The product $A_k\times B_k$ approximates the absolute communality of the k features (denoted as C_k) for the first component. Steps 3 and 4 will executate repeatedly if the sequence of absolute communalities appears an increasing trend. Otherwise, the stopping criterion is met.

Step 5. Use the latest subset of features to generate the multivariate classifier.

In this step, the first component obtained from the latest subset of features is outputted as the multivariate classifier of our inductive learning model.

5 Case Study

5.1 Dataset and Feature Evaluation

The data were taken from the German credit dataset preserved in the UCI repository [1]. There are 21 features (a_1~a_{21}) and 1000 instances in the dataset. These features are classified into three major categories: "Basic data", "Credit & financial grading", and "Job related rating". In the category of "Basic data", personal status & sex, present residence in years, telephone, age in years, housing status, purpose of loan, and other installment plans were included (as shown in Table 1). Personal status & sex, telephone, purpose of loan, and other installment plans were cardinal measurements with nominal scales. Present residence and age in years were measured with numerical scales, and housing status with an ordinal scale (1: rent, 2: own, 3: for free). In the category of "Credit & financial grading", except a_{18}, most features were measured with ordinal or numerical scales. In the category of "Job related rating", three features were measured with ordinal or cardinal scales.

To reduce computational complexity, numerical features are converted into discretized or nominal attributes. For comparison studies, the German credit dataset of preprocessed instances are saved into individual files so that they can be reused for building distinct decision models. All decision models were implemented in *C* and *Matlab* programming languages executed on a workstation with an Intel Core 2 dual 2.4 GHz processor. For verifying the accuracy of our classifier, 10-fold cross-validation was applied in the dataset. Namely, for each round of learning process, 9 folds (900 instances) were picked as the training data and the left (100 instances) were used as the test data. The training data were classified into four groups: High-credit, Good-credit, Normal-credit, and Bad-credit. Table 2 shows their definitions based on the features of a_8 and a_{18}. The data corresponding to a_8 were so objective that they are obtained from applicants' account. However, the data corresponding to a_{18} were more subjective since they were judged by the bank.

Table 1. Description of German credit dataset.

Feature	Scale
Basic data	
a_1: personal status & sex	cardinal (1~5)
a_2: present residence in years	numerical
a_3: telephone	cardinal (yes/no)
a_4: age in years	numerical
a_5: housing status	ordinal
a_6: purpose of loan	cardinal
a_7: other installment plans	cardinal
Credit & financial grading	
a_8: status of existing checking account	ordinal (1~4)
a_9: credit history	ordinal
a_{10}: credit amount	numerical
a_{11}: other debtors/guarantors	ordinal
a_{12}: number of existing credits at this bank	numerical
a_{13}: savings account/bonds	ordinal
a_{14}: installment rate in percentage of disposable income	ordinal
a_{15}: duration in months	numerical
a_{16}: number of people being liable to provide maintenance for	numerical
a_{17}: property ownership status	ordinal
a_{18}: credit final evaluation at this bank	cardinal (good/bad)
Job related rating	
a_{19}: present employment in years	ordinal
a_{20}: job attribute	cardinal
a_{21}: foreign worker	cardinal (yes/no)

Table 2. Four classes in German credit dataset.

	a_{18}=1 (good)	a_{18}=2 (bad)
a_8=1or 2 (higher or normal amount)	High-credit	Normal-credit
a_8=3 or 4 (lower amount or none)	Good-credit	Bad-credit

The scales and ranges of the remaining features varied from case to case. Take a_1 as an example. a_1=1 denoted male who is divorced or separated. a_1=2 denoted female who is divorced, separated, or married. a_1=3 denoted single male while 4 denoted male who is married or widowed. a_1=5 denoted single female. To measure the intra-diversity $ID(a_i,T)$ for the remaining features, the formula presented at the step 1 of Section 3 is applied 19 times. Table 3 displays the first six ($=\left|\sqrt{21}\right|+1$) features with the better *ID* measurements. The first four features belong to the "Credit & financial grading" category. As these measurements show, feature a_{16} is identified as the most representative variable of all and it truly makes sense that the number of people being liable to provide maintenance (guarantee) for the applicant is significant in assessing his credit

status. Property ownership status (i.e., a_{17}) and housing status (i.e., a_5) provide the real estate related information and are also identified as the import factors in assessing one's credit status. Installment rate in percentage of disposable income (i.e. a_{14}) has the *ID* measurement similar to a_{17} and a_5. Features a_2 and a_{19} are respectively numerical and ordinal variables whose *ID* measurements are significantly worst than the first four variables. We note that no cardinal feature is taken in the initial round of feature selection.

Table 3. The intra-diversity values.

Rank	Feature	*ID*(*x*,*T*)
1	a_{16}	17.55
2	a_{17}	45.50
3	a_5	46.30
4	a_{14}	46.75
5	a_2	56.81
6	a_{19}	80.73

5.2 Multivariate Classifier

After a small subset of effective features is ensured, the multivariate analysis was applied again to synthesize them. Suppose *n* features $a^{(1)}$, $a^{(2)}$, …, $a^{(n)}$ are inputted to the processing of PCA. And then the first component $y_1^{(n)}$ is outputted as the multivariate classifier. Alternatively and precisely, $y_1^{(n)}$ is denoted as $PCA[a^{(1)}, a^{(2)}, \ldots, a^{(n)}]$. Proceeding with the German credit data, $a^{(1)}=a_{16}$, $a^{(2)}=a_{17}$, $a^{(3)}=a_5$, $a^{(4)}=a_{14}$, $a^{(5)}=a_2$, and $a^{(6)}=a_{19}$ forms the base for the subsequent processing. Intuitively, $y_1^{(1)} = a_{16}$, $y_1^{(2)}=PCA[a_{16}, a_{17}]$, $y_1^{(3)}=PCA[a_{16}, a_{17}, a_5]$, and so on. Table 4 reveals the linear combination of variables and intra-diversity for each classifier. Notably, the absolute communality C_k 's keep a growing trend from $y_1^{(1)}$ to $y_1^{(3)}$ and thus $y_1^{(3)}=0.242a_{16}-0.667a_{17}-0.705a_5$ is concluded as the optimal classifier for the German credit data. The measurements of intra-diversity are also listed in the last column of Table 4 and they verify that $y_1^{(3)}$ owns the best evaluation of all. Since no cardinal feature is included in the classifier, the quantitative property can be completely reserved and then the computational results are meaningful.

5.3 Results

In this study, all learning processes were implemented in *Matlab* programming languages executing on a workstation with Pentium Dual-core CPU 2.00GHz processor. Besides, Weka 3.6 [2] data mining software in Java was also used to generate the related results about C4.5 decision tree for competitive studies. Our proposed model and C4.5 decision tree were trained 10 times for generating 10 results and then they were averaged to produce a single estimation.

To explore whether our model would support effective classification, analysis of variance (ANOVA) is applied to test if there was any significant difference between four classes. Table 5 shows that $y_1^{(2)}$ and $y_1^{(3)}$ have the better *F*-test and *p* values than others. In the last column of Table 5, the accumulated communalities measured from $y_1^{(3)}$ to $y_1^{(6)}$ indicate that they embrace sufficient population variances for classification task. However, for the sake of performance, $y_1^{(3)}$ is taken as the final classifier. So, statistical and quantitative analyses both support that $y_1^{(3)}$ is a fine and sound classifier in this experiment.

As to classification accuracy, two decision trees were built by C4.5 algorithm in Weka. In order to carry out a fair contest, the first decision tree only elected three features based on entropy and then constructs its classifier merely based on these features. Features a_6, a_9, and a_{11} were employed in the first decision tree. We denoted this model as DT(a_6, a_9, a_{11}). The second decision tree, denoted as DT(all), was built based on the scenario that all 19 features can be used. In a close look into test instances, the accuracies measured from our model and C4.5 decision tree in four classes (High-credit, Good-credit, Normal-credit, and Bad-credit) were listed in Table 6.

Table 4. Multivariate classifier.

Classifier	Linear combination of variables	C_k	*ID*(*y*,*T*)
$y_1^{(1)}$	a_{16}	0.0517	47.55
$y_1^{(2)}$	$-0.707a_{16}+0.707a_{17}$	0.2503	45.50
$y_1^{(3)}$	$0.242a_{16}-0.667a_{17}-0.705a_5$	0.4059	40.42
$y_1^{(4)}$	$0.158a_{16}-0.652a_{17}-0.686a_5-0.281a_{14}$	0.3549	45.50
$y_1^{(5)}$	$-0.154a_{16}+0.657a_{17}+0.642a_5+0.286a_{14}+0.225a_2$	0.3448	46.35
$y_1^{(6)}$	$0.182a_{16}-0.534a_{17}-0.503a_5-0.282a_{14}-0.357a_2+0.472a_{19}$	0.4065	60.20

Table 5. Statistical and quantitative analyses.

Classifier	*F*-test	*p* value	Accumulated communality
$y_1^{(2)}$	4.97	0.002	0.4967
$y_1^{(3)}$	4.97	0.002	0.8912
$y_1^{(4)}$	2.64	0.049	1.0153
$y_1^{(5)}$	1.69	0.168	1.2144
$y_1^{(6)}$	0.95	0.417	1.5618

The test set contains 100 instances for accuracy analysis. In the high-credit class with the largest data amount (39.5 instances in average for 10-fold cross-validation), the accuracy of our classifier is better than that of DT(all) and is nearly as good as that of DT(a_6, a_9, a_{11}). Note that there seems to have the losing accuracy on the bad-credit side in return for gaining accuracy on the high-credit side or vice versa for two decision tree models. High accuracy in identifying high-credit customer is so critical that their commercial contributions always dominate bank profits. Nevertheless, approving a potentially bad-credit customer is more costly than denying a good-credit customer. Our classifier successfully gains a better improvement in this respect. The customers accurately identified in the Good-credit class by the three models are respectively as 0.269×28.7(≅ 8), 0.086×28.7(≅ 2), and 0.414×28.7(≅ 12). For the normal-credit class, no customer is classified by three models. As to the bad-credit class, around 10, 7, and 10 customers are respectively identified by three models. The average amounts of instances corresponding to each class are listed in the last column of Table 6.

Table 6. The accuracies of the four classes.

Class	Accuracy			Size of class
	PCA-based	DT(a_6, a_9, a_{11})	DT(all)	
High-credit	0.844	0.896	0.584	39.5
Good-credit	0.269	0.086	0.414	28.7
Normal-credit	0	0	0.091	4.6
Bad-credit	0.352	0.241	0.370	27.2
Total				100.00

Finally, the amounts of data entries involved in the training processes of three models are counted for efficiency analysis. For each try in the 10-fold cross-validation experiment, 19 features and 900 instances have to be taken in feature evaluation. Namely, at least 19×900=17100 data entries must be involved for each model. Besides the root node, C4.5 algorithm executes several feature evaluation rounds in order to select the proper features for the sequent decision nodes. This is why Table 7 shows our PCA-based model imposes on less data than C4.5. Table 7 verifies that time efficiency is successfully achieved by our model.

Table 7. Efficiency analysis.

Model	Number of data entries	Ratio
PCA-based	17100	1
DT(a_6, a_9, a_{11})	21255	1.25
DT(all)	86890	5.08

6 Conclusion

Empirically, in a real data analysis, multiple class problems are frequently encountered, and a classifier which explains multiple classes simultaneously would be valuable and would exhibit higher readability. Many application domains (e.g., environmental science, actuarial science, biostatistics, population ecology, psychometrics, and quality control) highlight the usefulness in solving multi-class classification or optimization problems. When more alternative financial plans are launched for serving different customers, the decision makers of financial institutions have to promote many adequate plans to satisfy various customers' requirements. It needs an efficient customer identification mechanism or a powerful classifier with acceptable prediction accuracy to solve such multi-class classification or optimization problems. This paper proposes an inductive learning model which captures the effective information after a more careful examination. A sounder classifier based on multivariate analysis is achieved by this model. Besides the effectiveness and efficiency verified by our experimental results, our model has flexibilities in three aspects. The first aspect is about feature evaluation method. In accordance with the practical need, data analyzers can replace the entropy-based measures with other methods such as Gini Index [17], goodness, or impurity. The second aspect regards computational performance. Our model can speed up the learning process by considering fewer features in case that analysis efficiency is more crucial than accuracy to the decision making. Finally, our model can cooperate with *k*-NN algorithm in retrieving a collection of similarity to every new query object for accurate reasoning.

References:

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/

[2] http://www.cs.waikato.ac.nz/ml/weka/

[3] L. Becchetti, C. Castillo, D. Donato, R. Baeza-Yates, S. Leonardi, Link analysis for Web spam detection, *ACM Transactions on the Web* (TWEB), Vol.2, No.1, 2008.

[4] T. Bellotti, J. Crook, Support vector machines for credit scoring and discovery of significant features, *Expert Systems with Applications*, Vol.36, No.2, 2009, pp. 3302-3308.

[5] R.J. Bolton, D.J Hand, Statistical fraud detection: a review, *Statistical Science*, Vol.17, No.3, 2002, pp. 235–255.

[6] C. Castillo, D. Donato, A. Gionis, V. Murdock, F. Silvestri, Know your neighbors: web spam detection using the web topology, *Annual ACM Conference on Research and Development in Information Retrieval*, Amsterdam, Netherlands, 2007, pp. 423-430.

[7] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, *ACM Computing Surveys* (CSUR), Vol. 41, No.3, 2009.

[8] L. Chen, Q. Ye, Y. Li, Research on GA-based bank customer's credit evaluation, *Computer Engineering*, Vol.32, No.3, 2007, pp. 70-72.

[9] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, *Artificial Intelligence in Medicine*, Vol.34, No.2, 2005, pp. 113-127.

[10] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, *Journal of Bioinformatics and Computational Biology*, Vol.3, No.2, 2005, pp. 185–205.

[11] V. Estivill-Castro, I. Lee, Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data, *Proceedings of the 6th International Conference on Geocomputation*, 2001.

[12] J. A. Etzel, V. Gazzola, C. Keysers, Testing simulation theory with cross-modal multivariate classification of fMRI data. *PLoS ONE*, Vol.3, N.11, 2008.

[13] Y. Feng, Z. Wu, X. Zhou, Z. Zhou, W. Fan, Knowledge discovery in traditional Chinese medicine: State of the art and perspectives. *Artificial Intelligence in Medicine*, Vol.38, No.3, 2006, pp. 219-236.

[14] K. Fukunaga, Introduction to Statistical Pattern Recognition (2nd ed.). Academic Press, 1990.

[15] A. Genkin, D.D. Lewis, D. Madigan, Large-scale Bayesian logistic regression for text categorization. *Technometrics*, Vol.49, No.3, 2007, pp. 291-304.

[16] X. Hao, W. Deng-sheng, X. Yang-Qun, Study on enterprise credit evaluation based on PCA/FCM, *Technology Economics*, 3, 2007.

[17] A.J. Hawyard, Mathematics and Politics (New York: Macmillan Company, 1965), as cited in a 1989 computer program by Tom Finholt, Department of Social and Decision Sciences, Carnegie Mellon University.

[18] A. Khashman, Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, Vol.37, No.9, 2010, pp. 6233-6239.

[19] H.-W. Kim, H. C. Chan, S. Gupta, Value-based adoption of mobile internet: An empirical investigation, *Decision Support Systems*, Vol.43, No.1, 2007, pp. 111-126.

[20] C. Orsenigo, C.Vercellis, Multivariate classification trees based on minimum features discrete support vector machines, *IMA Journal of Management Mathematics*, Vol.14, No.3, 2003, pp. 221-234.

[21] C. Piao, J.An, M. Fang, Study on credit evaluation model and algorithm for C2C E-Commerce, *IEEE International Conference on e-Business Engineering*, Hong Kong, 2007, pp. 392-395.

[22] X. Qi, B.D. Davison, Web page classification: Features and algorithms. *ACM Computing Surveys* (CSUR), Vol.41, No.2, 2009.

[23] M. Šušteršič, D. Mramor, J. Zupan, (). Consumer credit scoring models with limited data. *Expert Systems with Applications*, Vol.36, No.3, 2009, pp. 4736-4744.

[24] C.F. Tsai, Feature selection in bankruptcy prediction, *Knowledge-Based Systems*, Vol.22, No.2, 2009, pp. 120-127.

[25] C. Wu, H. Xia, Study of personal credit evaluation under C2C environment based on support vector machines ensemble, *International Conference on Management Science and Engineering*, pp. 25-31, CA: Long Beach, 2008.

[26] L. Yu, W. Yue, S. Wang, K.K. Lai, Support vector machine based multiagent ensemble learning for credit risk evaluation, *Expert Systems with Applications*, Vol.37, No.2, 2010, pp. 1351-1360.

[27] H. Zhao, A multi-objective genetic programming approach to developing Pareto optimal decision trees, *Decision Support Systems*, Vol.43, No.3, 2007, pp. 809-826.

[28] M. Zucknick, S. Richardson, E.A. Stronach, Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods, *Statistical Applications in Genetics and Molecular Biology*, Vol.7, No.1, 2008.