

國立交通大學
工業工程與管理學系
碩士論文

建構複合式信用評等模型

Constructing Hybrid Credit Scoring Model



研究生：游翔百

指導教授：唐麗英 博士

中華民國九十三年七月

建構複合式信用評等模型

Constructing Hybrid Credit Scoring Model

研究生：游翔百

Student : Hsiang-Pai Yu

指導教授：唐麗英 博士

Advisor : Lee-Ing Tong

國立交通大學

工業工程與管理學系

碩士論文

A Thesis

Submitted to Department of Industrial Engineering and Management

College of Management

National Chiao Tung University

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

In

Industrial Engineering

July 2004

Hsin-Chu, Taiwan

Republic of China

中華民國九十三年七月

建構複合式信用評等模型

研究生：游翔百

指導教授：唐麗英

國立交通大學工業工程與管理學系碩士班

摘要

風險評估及信用評等是金融機構用以評量借款企業償債能力的重要依據，然而身處目前經濟不景氣的環境下，逐漸身高的逾放比率使得越來越多的金融機構必須檢討其現有信用評等模式的缺失，以對貸款企業做出更正確有效的放款決策。現有之中、外文獻雖發展出許多信用評等模式來探討此問題，一般說來以類神經網路架構出的信用評等模型分類正確率表現較傳統統計方法建構出之模型為佳。但基於類神經網路(Artificial Neural Networks: ANN)理論上的不足，使得類神經網路架構出之信用評等模型解釋能力不佳，在實務層面上難以使用。因而本研究乃針對台灣金融機構之中小企業借款者，發展出一套複合式信用評等模型，此模型流程首先建立分類迴歸樹(Classification and Regression Tree: CART)，然後再將分類迴歸樹的預測結果及事後機率作為後續的類神經網路的輸入變數，藉此來增加整體複合式信用評等模型的分類正確率；此外，藉由使用分類迴歸樹來鑑別具有顯著影響的變數，增加整體複合式信用評等模型的模型解釋能力。

同時，本研究也廣泛比較現存的信用評等模型預測能力的差異，分別利用了線性判別分析(Linear Discriminant Analysis: LDA)、曲線判別分析(Quadratic Discriminant Analysis: QDA)、羅吉斯迴歸(Logistic Regression: LR)、機率類神經網路(Probabilistic Neural Network: PNN)、倒傳遞網路(Back Propagation Neural Network: BPN)、一般迴歸神經網路(General Regression Neural Network: GRNN)、自組性演算法(Group Method of Data Handling: GMDH)、K最近鄰居法(K-Nearest Neighbor: KNN)及學習向量量化網路(Learning Vector Quantization Neural Network: LVQ)等不同的信用評等模型，透過台灣某金融機構所提供中小企業借款者的實際歷史資料，驗證了本研究所提出之複合式信用評等模型確實有效可行。

【關鍵詞】：信用評等、分類迴歸樹、複合式模型、類神經網路

Constructing Hybrid Credit Scoring Model

Student : Hsiang-Pai Yu

Adviser : Lee-Ing Tong

Department of Industrial Engineering and Management
National Chiao Tung University

Abstract

Credit scoring is an essential task for banks and loan companies in the last few decades. The demand of developing a credit scoring model with reliable accuracy has become an urgent issue. Among many studies of credit scoring, artificial neural network (ANN) is a promising technique to achieve high accuracy of classification compared to existing conventional techniques. However, the poor explanation power makes ANN difficult to produce interpretable result. This drawback also decreases the power of ANN applied in practical problems. The objective of this study is to propose a hybrid credit scoring model which is combined with CART and other algorithms to enhance the accuracy of credit scoring model, and increase the interpretable capability as well. Financial loan companies can employ this study when establishing their credit scoring models.

.

Key Words: Hybrid model, Artificial Neural Network (ANN), Credit Scoring, CART

誌 謝

這本論文的完成，有著許多人的幫助與心血。最值得感謝的人是我的指導教授唐麗英博士，若沒有她辛勤的審閱我的論文，這本論文的可讀性一定慘不忍睹。在這兩年的交大研究生生活，唐麗英老師啟發了我許多為人處事以及作學問的正確態度，同時也讓我對自己未來的人生規劃，有著更清晰的藍圖。同時，老師對我的諸多提攜，總是提供許多機會讓我去嘗試，像是論文比賽、出國研討會發表論文等，使我在這兩年間大有成長，這都得感謝唐麗英老師。

同時感謝口試委員梁高榮教授、王春和教授和計畫書審查委員李慶恩教授提供諸多建議，使得這本論文更臻完善。

碩士班兩年非常充實而有趣，感謝實驗室夥伴這兩年的陪伴與協助，千慧學姐、民祥、俊誠、文傑、宏志、冠人、政勳、忠佐、盛全等，我衷心感謝大家，點點滴滴的回憶都將銘記我心。同時，MB606 實驗室的楓凱、淙亮、英泰、士凱等；MB604 實驗室的渙群、石隆；MB002 實驗室的盈月、仁耀；管科所的詠涵、佩雙、慧菁等謝謝你們一路的陪伴，狂歡生日會、螢火蟲夜遊、熬夜寫報告、俊誠家夜烤等等，這些回憶因為有你們的參與，而令人難忘。

還得感謝我的女朋友馨平，這兩年她的對我的忍耐與扶持，是我能努力完成此本論文的最大因素，每每碰到挫折與疲累，只有她會一再鼓勵與支持我，讓我持續下去，說她是我精神上的支柱一點也不為過。

最後要感謝的是我的父母，這兩年我不常回家，我的父母也毫無怨言，只有辛勤的噓寒問暖，沒有他們，我絕不可能念到碩士學位。僅以此文，向我的父母親及曾幫助我的許多人表達心中最誠摯的感謝。

游翔百 謹誌於

交通大學工業工程與管理研究所

2004年7月15日

Contents

中文摘要.....	I
英文摘要.....	II
誌謝.....	
Contents.....	IV
List of Figures.....	
List of Tables.....	VIII
Chapter 1 Introduction.	1
1.1 Background and Motivation.....	1
1.2 Research Objective.....	3
1.3 Organization.....	4
Chapter 2 Literature Review.	5
2.1 Contemporary credit scoring system of bank loaning.....	5
2.1.1 The origin and development of credit scoring.....	5
2.2 Discriminant analysis.....	6
2.3 Artificial Neural Networks.....	7
2.4 Other nonparametric methods.....	11

2.4.1 Classification and regression tree (CART)	11
2.4.1.1 Classification Tree Methodology.....	11
2.4.1.2 Tree impurity function.....	13
2.4.1.3 Tree growing methodology.....	13
2.4.1.4 Tree Pruning.....	14
2.4.2 Group Methods of Data Handling (GMDH).....	16
2.4.3 General Regression Neural Network (GRNN).	18
Chapter 3 The Proposed Hybrid Model Approach.	20
3.1 Model Evaluation Criterion.	20
3.2 Procedure of Constructing Hybrid Credit Scoring Model.	21
3.3 Procedure of Establishing Prediction Model of Default Period.	27
Chapter 4 Illustrative Examples.	31
4.1 Description of Sample Data.	31
4.2 Perform CART.	32
4.3 Record CART's Split Variables and Predictive Outcomes.	35
4.4 Use recorded variables and predictive outcomes as input variables of following model.	35
4.5 Compare The Accuracy of each Hybrid Model and Choose the	

Best Hybrid Credit Scoring Model. 46

4.6 Establish Prediction Model of Default Period. 46

4.7 Further Comparison of Hybrid Model. 47

Chapter 5 Concluding Remarks 50

References.....52

Appendix.....55



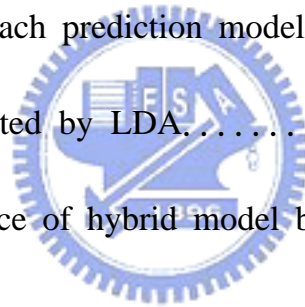
List of Figures

Figure 1. Logit tree diagram of credit scoring model.....	2
Figure 2. The traditional scope of credit scoring model.....	3
Figure 3. Three layer back propagation neural network (BPN).....	8
Figure 4. Example of classification tree.....	12
Figure 5. Geometric Viewpoint of CART.....	13
Figure 6. Multilayered structure of GMDH with five inputs and selected nodes	18
Figure 7. Flowchart of the proposed hybrid model.....	21
Figure 8. Hybrid BPN credit scoring model.....	27
Figure 9. Flowchart of default period prediction model.....	29
Figure 10. Original BPN model accuracy.....	38
Figure 11. Hybrid BPN produced by Cart_1.....	39
Figure 12. Hybrid BPN produced by Cart_2.....	39
Figure 13. Hybrid BPN produced by Cart_3.....	40
Figure 14. Hybrid BPN produced by Cart_4.....	40
Figure 15. Hybrid BPN produced by Cart_5.....	41
Figure 16. Hybrid BPN produced by Cart_6.....	41

List of Tables

Table 1.	Financial strength ratings of S&P corp.	6
Table 2.	Merits and demerits of artificial neural networks by Vellido.	10
Table 3.	Variable Description.	31
Table 4.	CART candidate models.	33
Table 5.	CART`s split variables.	33
Table 6.	Comparison of original credit scoring models.	34
Table 7.	Input Variables of following hybrid models.	35
Table 8.	Hybrid LDA Performance.	36
Table 9.	Original LDA Performance.	36
Table 10.	Hybrid QDA Performance.	36
Table 11.	Original QDA Performance.	36
Table 12.	Hybrid LR Performance.	37
Table 13.	Original LR Performance.	37
Table 14.	Hybrid PNN Performance.	42
Table 15.	Original PNN Performance.	42
Table 16.	Hybrid GRNN Performance.	43
Table 17.	Original GRNN Performance.	43
Table 18.	Hybrid GMDH Performance.	43

Table 19.	Original GMDH Performance.....	43
Table 20.	Hybrid KNN Performance.....	44
Table 21.	Original KNN Performance.....	44
Table 20.	Hybrid KNN Performance.....	44
Table 21.	Original KNN Performance.....	44
Table 22.	Hybrid LVQ Performance.....	45
Table 23.	Original LVQ Performance.....	45
Table 24.	Best Hybrid Credit Scoring Model.....	46
Table 25.	The MSE of each prediction model	45
Table 26.	Variables selected by LDA.....	48
Table 27.	The performance of hybrid model based on LDA.....	48



Chapter1 Introduction

1.1 Background and Motivation

With rapid growth of the credit industry in last few decades, credit evaluation of loan applicants becomes an important issue not only because the urgent demand from bankers, but also due to the pressure of cash flow and collections [5]. The conventional credit scoring or credit evaluation models simply classify loan applicants into two categories: “Good Loaner” and “Bad Loaner” according to some financial studies. Credit decision-makers can use the result of credit evaluation to make the right judgment and minimize bad loan risk. As a result, credit evaluation received more attention by bankers and a trustworthy credit scoring model became an urgent issue. With a sizable loan portfolio, even slight improvement in the accuracy of credit evaluation can reduce the creditors’ risk and translate the accuracy improvement considerably into future savings, cost reduction, faster credit evaluation, and closer monitoring of existing accounts. [5].

In the past, credit scoring was evaluated by creditor analysts. Due to the sharp growth of credit industry, the workload of credit analysts has exceeded its capacity. As a result, finding new automated ways of credit evaluation has become a forthcoming trend.

In addition, the risk of potential bad debts is also another critical issue. The depression of financial market made loan applicants of mid-size companies endure greater default pressure than they had in the past. Therefore, loan companies necessitate an accurate credit scoring model urgently to classify loan applicants to alleviate potential loss of bad debt. The percentage of bad loans increases rapidly, credit analysts are looking for strict and objective measures to evaluate loan applicants. All agendas discussed above can be shown in Fig.1.

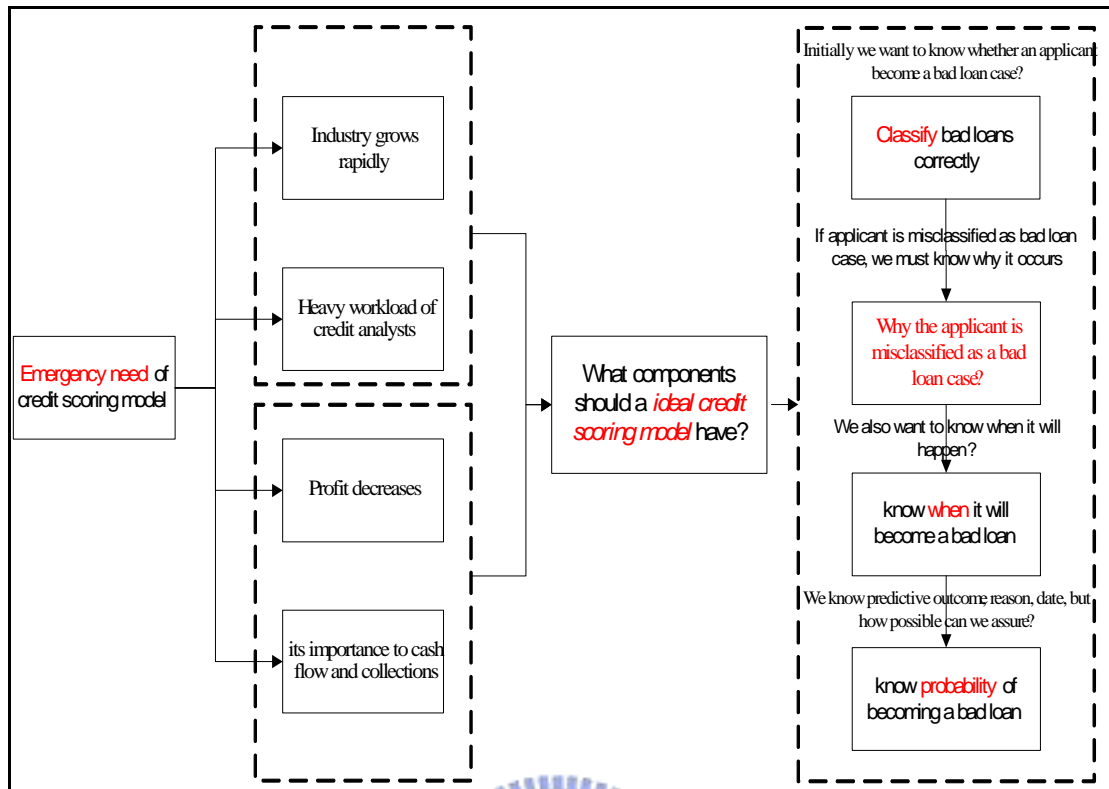


Figure 1. Logit tree diagram of credit scoring model

Many studies on credit scoring model were mainly focused on improving the prediction through various methods such as decision tree (CART) [3], logistic regression (LR) [6,18,20], linear discriminant analysis (LDA) [2,3], k-nearest neighbor (KNN)[20], and artificial neural networks (ANN) [3,6,9,11,12,13,14,20]. In other words, previous studies elevated on one dimension only--either on classification accuracy or on interpretable capability. Although accuracy or interpretable capability are two major criteria for assessing a credit scoring model, optimizing two major criteria simultaneously are challengeable. That is, pursuing promising classification accuracy and seeking interpretable capability lie on a “trade-off” relation as shown in Fig.2.

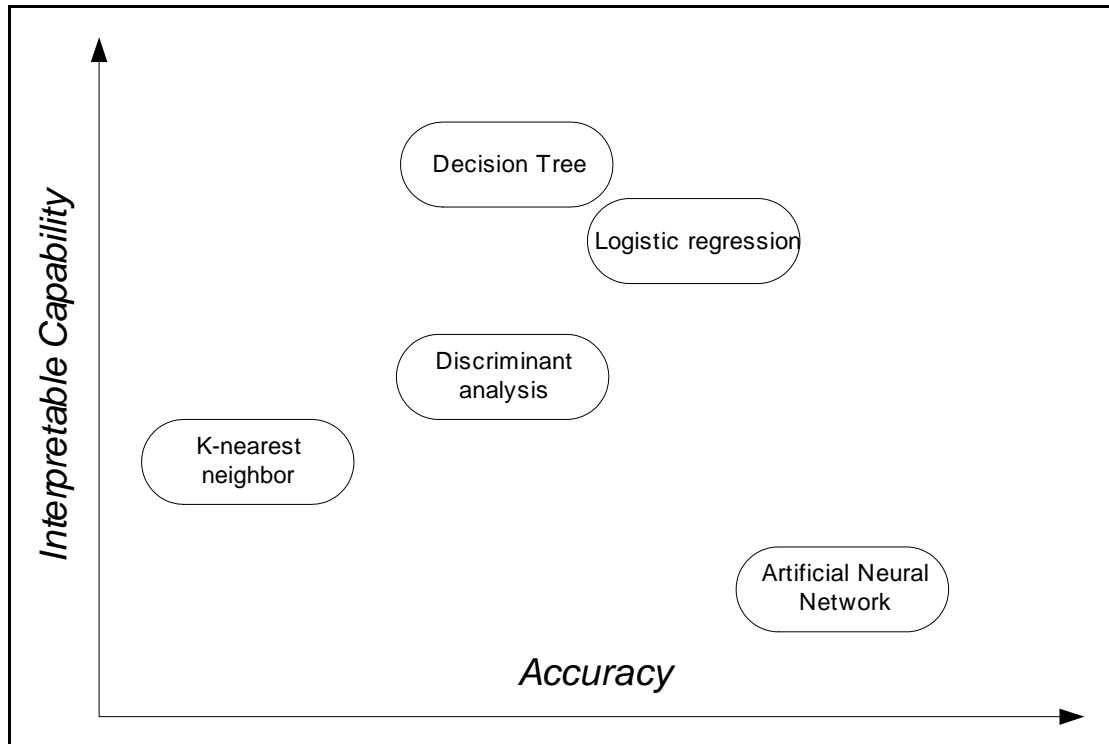


Figure 2. Traditional scope of credit scoring model

ANN has proved its capability of classification and prediction accuracy in constructing credit scoring model. Many studies indicated that ANN has superior or even dominate accuracy as compared to many conventional statistical classification methods. However, ANN still has some drawbacks such as “black box procedure”, “lack of explanation”, “complex network design”, “lack of feature selection” etc. Among these drawbacks, failing to interpret the classification results is the most controversial problem of ANN. Decision makers occasionally think it is hard to utilize ANN’s results in practice because of the above drawbacks.

1.2 Research Objective

The objectives of this study can be summarized as follows:

1. Construct a hybrid credit scoring model with superior classification accuracy and interpretable capability than existing credit scoring models.
2. Develop a simple feature selection method to enhance the capability of interpretation.

This study utilized decision tree, which adopting classification and regression tree (CART) [4] algorithm, as a feature selection method to choose significant input variables. The chosen variables are then used as inputs for ANN and enhance the total predictive accuracy. In other words, CART can be regarded as a “guide” to construct the credit scoring model, followed by the ANN model to realize not only the influential input variables but also CART’s own classification results. As a result, we may expect the following ANN model or other complex algorithms can learn more accurately and quickly on account of good guide “CART”. Several data mining algorithms were utilized to replace ANN to obtain a best hybrid credit scoring model.

1.3 Organization

The rest of this study is organized as follows. Chapter 2 reviewed the related fundamentals of credit scoring model. Chapter 3 described the proposed method and introduced the model evaluation criterion in detail. Chapter 4 presented the illustrative examples and compared the proposed hybrid approach with existing credit scoring models to demonstrate the effectiveness of the proposed hybrid models. Chapter 5 summarized the result of the study and further research direction.

Chapter 2 Literature Review

In this chapter, related fundamentals and studies are reviewed. Section 2.1 is a brief introduction of the existing credit scoring system of bank loaning. Section 2.2 discusses reviews the pros and cons of discriminant analysis (DA) used in credit scoring in the past. Section 2.3 presents the artificial neural network (ANN) approach and discusses the related issues of pros and cons. Section 2.4 reviews nonparametric methods and discusses their drawbacks.

2.1 Contemporary credit scoring system of bank loaning

The conventional procedure of constructing a credit scoring of bank loaning is to evaluate the corresponding credit factors such as financial variables and non-financial variables of a company, and credit analysts aggregate the evaluation scores from the credit factors and make the decision. Obviously this procedure lacks objectivity, and credit analysts can easily be misled because of insufficient priori knowledge. Moreover, the result of this credit scoring model may be easily dominated by few key analysts who own the power. Therefore, many studies dedicated to develop a quantitative credit scoring model to avoid shortcomings of the conventional models.

2.1.1 The origin and development of credit scoring

There are over twenty renowned credit scoring companies in the world. “Moody’s”, “Standard & Poor’s (S&P)”, “Fitch IBCA” are three most prominent companies among them and their credit assessing results are widely adopted as external credit scoring models by banks in the whole world. Table 1 presents the rating standard and the corresponding financial strength of companies by S&P. The rating standard can be used as a primary reference for external credit scoring.

Table 1 Financial strength ratings of S&P corp.

	Rank	Corresponding meanings
Safety	AAA	Extremely Strong
	AA	Very Strong
	A	Strong
	BBB	Good
Weak	BB	Marginal
	B	Weak
	CCC	Very Weak
	CC	Extremely Weak
	R	Under Regularly Supervision
	NR	Not Rated

2.2 Discriminant Analysis

Linear discriminant analysis (LDA) [17] is the first classification algorithm applied in credit scoring. LDA has been the most commonly used statistical technique in constructing classification model because of its simplicity and popularity. LDA attempts to find a linear combination of predictor variables to classify objects into various groups. Discriminant analysis is designed to maximize the ratio

$$\lambda = \frac{\gamma^T B \gamma}{\gamma^T W \gamma},$$

where γ is a $p \times 1$ vector of weights, B and W represent the between-groups and within-group sum of squares for the discriminant function ξ , respectively. The discriminant function is given by

$$\xi = X^T \gamma,$$

where X is a $p \times 1$ random vector of p variables. Analytically, the objective of DA is to identify the weights γ such that the ratio λ is maximized.

Altman [2] collected 33 bankrupt companies and 33 contrary healthy companies to construct a LDA credit scoring model. He found that the linear discriminant credit

scoring model performed very well, especially in short time period. Lee *et al.* [13] integrated the BPN and LDA approaches to obtain a hybrid credit scoring model and showed that the proposed hybrid approach converges much faster than the conventional BPN model. Moreover, his results indicated that the credit scoring accuracies of the hybrid model outperforms the original BPN, LDA and logistic regression (LR) approaches. A similar study presided by Lee *et al.* [12] also considered the hybrid neural network models for bankruptcy predictions. Their hybrid methodology contains multiple discriminant analysis (MDA)-assisted neural network, the ID3-assisted neural network operated with the input variables selected by the MDA method and ID3, respectively. They concluded that the hybrid neural network models are very promising for bankruptcy prediction in terms of predictive accuracy and adaptability. Markham and Ragsdale [14] observed that combining the predictions of a well-known statistical tool with one of ANN techniques may provide more accurate prediction results than either individual techniques used alone. They utilized Mahalanobis distance measure (MDM) as inputs of ANN and showed that the hybrid methodology can significantly reduce the average misclassification rate.

However, the utilization of LDA in constructing the credit scoring model has received many criticisms because of its theoretical assumptions, such as data must possess a multivariate normal distribution, and the covariance matrices of good loan and bad loan classes must be equal, are frequently violated in real-world data [6,10,20]. Although quadratic discriminant analysis (QDA) can alleviate some drawbacks of LDA, QDA does not perform better than LDA as expected [10,17].

2.3 Artificial Neural Networks

Many researches explored the capability of ANN applied in business problems such as credit scoring or bankruptcy prediction. ANN can learn complex non-linear structure of datasets or can approximate many continuous functions accurately.

Besides, ANN does not require any priori assumptions about data distribution. A large number of researches and surveys have proven that ANN is a suitable and outstanding technique on extensive business applications [6,9,11,12,13,14,15,18,19,20].

ANN generally consists of three layers: an input layer, a hidden layer, and an output layer. Each layer is interconnected by a number of processing units called “neurons” or “nodes”. Each unit represents a computation device and it transforms an input to an output by means of some pre-specified function. Each link is assigned a numerical value representing the weight of connection. Input nodes receive input signals and aggregate information into hidden layer nodes, and the hidden layer nodes transform the aggregate information into desired targets in output layer nodes by some pre-specified activation function. ANN iteratively adjusts network weights in order to produce desired output as closer as possible. The value of network weight is determined by inputs and outputs of the training dataset through learning algorithms. The objective of ANN is to find a set of appropriate network weights under different network topologies and predict or classify observations accurately. Figure 3 shows a brief presentation of ANN.

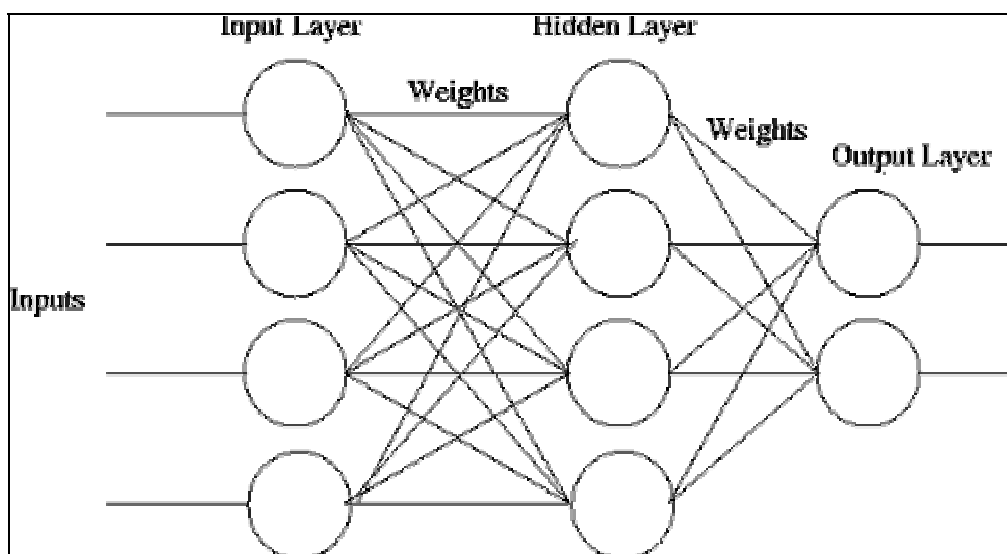


Figure 3 Three-layer back propagation neural network (BPN)

Piramuthu et al. [16] used feature construction to improve the performance of ANN and assessed their proposed methodology using Belgian bankruptcy data. Their study concluded that the feature construction improves the searching procedure through the solution-space and increases the average information content of each feature which is used as input to BPN.

Olmeda and Fernandez [15] compared the accuracy of several classifiers on the problem of bankruptcy prediction. They concluded that ANN provided the best results compared to logistic regression, DA, C4.5 and multivariate adaptive regression spline (MARs).

Tam and Kiang [18] compared a number of well-known classifiers such as DA, logistic regression (LR), k nearest neighbor (KNN), ID3, and BPN applied in bank failure predictions. Their results indicated that modified ANN with given prior probabilities and misclassification costs was a promising method of evaluating bank conditions in terms of predictive accuracy, adaptability, and robustness.

West [20] investigated the accuracy of credit scoring model constructed using five neural network approaches: multilayer perceptron, mixture-of-experts, radial basis function, learning vector quantization, and fuzzy adaptive resonance. The results are benchmarked against some traditional methods including DA, LR, KNN, kernel density estimation, and CART. His study concluded that BPN may not be the most accurate model and logistic regression is found to be the most accurate traditional method for building a credit scoring model except for ANN approaches.

Vellido et al. [19] has surveyed extensively and found that 74 out of 93 papers relied on using the back propagation neural network (BPN), a few others utilized learning vector quantization (LVQ), radial basis function (RBF), self-organization map (SOM), etc. With respect to credit scoring related researches, BPN is the most

widespread model and is often used as a benchmarking approach for other models.

Zhang [21] have shown in full details that among all controversial criteria disputed in ANN for classification and summarized, that two of the most important developments in ANN classification were the studies of hybrid neural model and feature selection.

One of the disadvantage of ANN is the poor explanatory capability which is referred to as “black box” problem. Because ANN is unable to identify the influential variables or the relevant variables, the result of ANN model may be difficult to achieve rational explanations. Another disadvantage of ANN is that ANN is lack of formal explanation on neural network architecture, that is, there is no formal procedure either to select network topology or to decide network architecture. Vellido [19] indicated that the rule of thumb is the most popular way to select the network topology or decide network architecture. Some researchers such as Glorfeld and Hardgrave [9], Piramuthu *et al.* [16] endeavored to develop some modification or rules of existing algorithms, but could not obtain satisfactory results.

Table 2. Merits and demerits of artificial neural networks by Vellido [19]

Merits of artificial neural network	Demerits of artificial neural network
<ul style="list-style-type: none"> • Able to learn any complex nonlinear mapping or approximate any continuous function • As non-parametric methods, NN do not make any priori assumptions about the distribution of data or input-output mapping function • NN are very flexible with respect to incomplete, missing and noisy data, NN are “fault tolerant” • Neural network models can be easily updated / are suitable for dynamic environments. 	<ul style="list-style-type: none"> • Lack theoretical background concerning explanatory capabilities and results in “black boxes” • The selection of the network topology and its parameters lack theoretical background, it is still a “trial and error” matter. • Training process of NN is very time-consuming. • Neural network can overfit the training data and lose generalization capability.

2.4 Other Nonparametric Methods

Nonparametric methods such as LR and CART can be applied in constructing credit scoring model. However, a number of comparative studies indicated that these methods perform well only in specific environment. West [20] also pointed that nonparametric methods do not provide satisfactory outcomes in many studies.

However, predictive accuracy is not the only concerned perspective in constructing credit scoring model. Decision tree, K-nearest neighbor (KNN) or other nonparametric methods can also be used as preprocessing mechanisms to enhance the performance of ANN. Vellido [19], Lee et al. [12], Lee et al. [13], Markham and Ragsdale [14] explored the performance of hybrid model and their results showed that the hybrid model performed better than the original ANN methods in respect of predictive accuracy and speed of convergence.

2.4.1 Classification and Regression Tree (CART)

CART [4] is a decision tree method for analyzing categorical data as a function of continuous or categorical explanatory variables. CART uses a set of training samples to grow a classification tree and prune a tree, and finally utilizes a set of testing samples to determine the right size tree which has the lowest misclassification cost.

2.4.1.1 Classification Tree Methodology

A classification tree T for a categorical variable is constructed by employing recursive partitioning the training samples into two different subsets. The objective is to find the appropriate explanatory variables that can split the training samples as correct as possible according to some pre-specific splitting criteria. The subsamples are called leaf nodes or nodes. The entire original training samples are noted as root node t_l of the tree. Similarly, the descendent nodes are abbreviated as t_L for the left subsamples. Subsamples which are not split further are called the terminal nodes. Graphically, the nodes and splitting rules denoted under each node are depicted in

Fig.4.

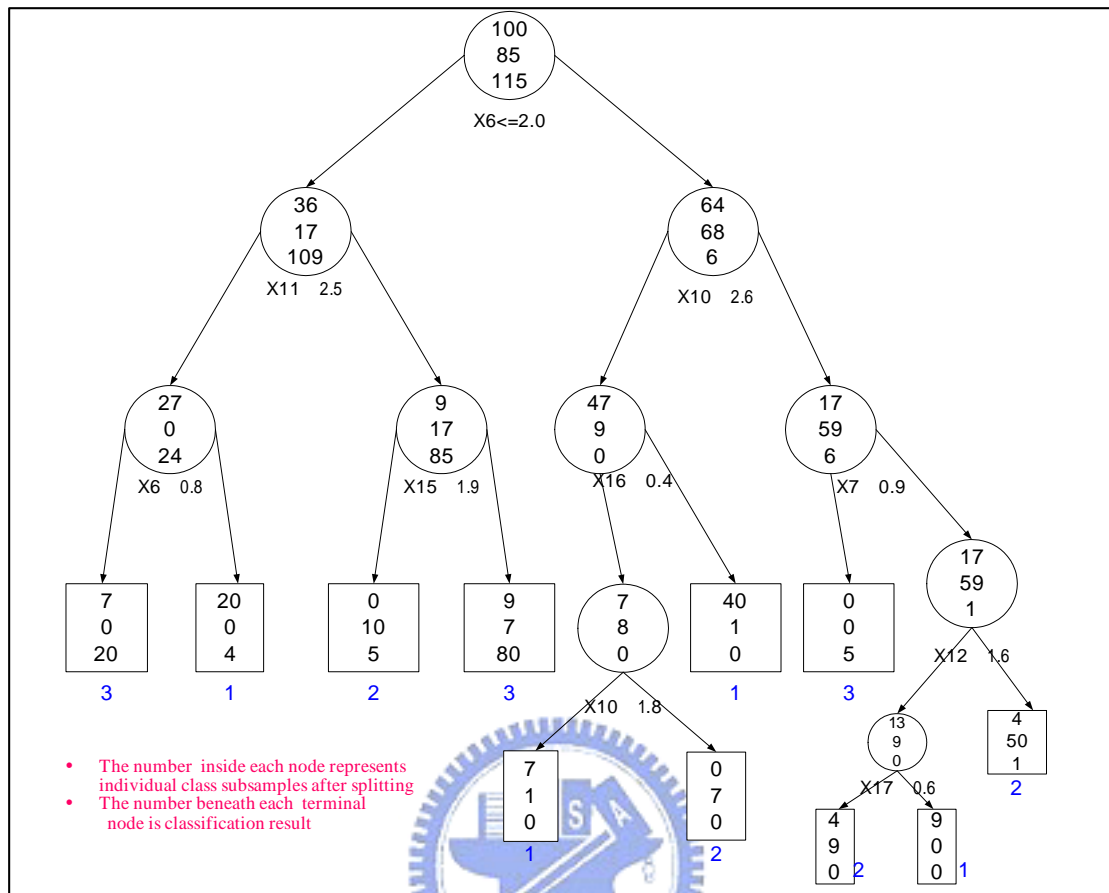


Figure 4 Example of classification tree

The splitting criterion of CART is to split the training sample into two subsamples. Each of the subsamples contains only cases from one category. In this case, the split decreases the most impurity of parent node, in other words, the tree can be thought of as a “partitioning hyperplane into rectangle” such that the populations within each rectangle become more and more homogeneous. Fig.5 depicts the case.

The impurity measure $i(t)$ of node t is defined as $i(t) = (p(1|t), p(2|t), p(J|t))$. The node impurity is the largest when all classes are equally mixed together in node, and it becomes the smallest in the case where the node contains only one class. Our goal is to decide the best split which decreases the impurity as much as possible.

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

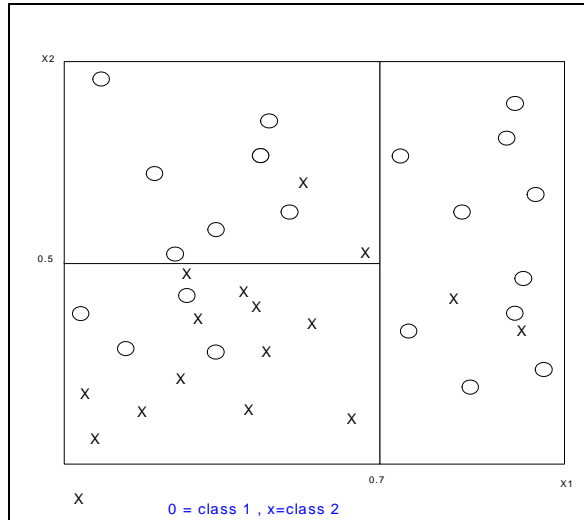


Figure 5 Geometric Viewpoint of CART

2.4.1.2 Tree impurity function

A tree impurity $I(T)$ can be defined as follows:

$$I(T) = \sum_{t \in \bar{T}} I(t) = \sum_{t \in \bar{T}} i(t)p(t)$$

Maximizing the decrease in tree impurity $I(T)$ by splitting the number “ s ” on node t is given as follows:

$$\begin{aligned} \Delta I(s, t) &= I(t) - I(t_L) - I(t_R) \\ &= \Delta i(s, t)p(t) \end{aligned}$$

2.4.1.3 Tree growing methodology

There are five steps for employing tree growing methodology:

1. Decide impurity function
2. Grow tree by maximizing tree impurity decrease until the tree size become as large as possible.
3. Get the best tree by pruning structure.
4. Use the proper “ estimation method ” to get estimator of tree classifier.
5. Interpretation results.

2.4.1.4 Tree pruning

In CART algorithm, it adopts “Minimal cost-complexity pruning” to prune the tree:

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}|$$

$$R_\alpha(T(\alpha)) = \min_{T \leq T_{\max}} R_\alpha(T)$$

Above formula implies that α can be thought as the complexity cost per terminal node and R is an linear combination of the total misclassification error $R(T)$ and its complexity cost $\alpha |\tilde{T}|$ of subtree T . As the penalty cost of complexity per terminal node increases, the minimizing subtrees $T(\alpha)$ will have fewer terminal nodes. When α is large enough, the subtree $T(\alpha)$ will eventually consist of the root only, and the tree T_{\max} will be completely pruned.

The pruning outcomes are expected to be:

$$T_{\max} > T_1 > T_2 > \dots > \{t\}$$

However, the above outcomes are hard to achieve. Neither $T_1 > T_2$, nor T_2 is necessarily pruned from previous subtree T_1 . Direct search through all possible subtrees to find $R_\alpha(T)$ is computationally expensive. As a result, Breiman [3] used “Weakest-Link cutting” for any non-terminal node t of T_{\max} which appears $R(t) > R(T_t)$. Actually t can be thought as the survival node of pruned tree after removing branch tree T_t .

$$R_\alpha(\{t\}) = R(t) + \alpha \times 1$$

$$R_\alpha(\tilde{T}_t) = R(T_t) + \alpha \times |\tilde{T}_t|$$

The subtree t means the subtree contains only one terminal node t , and its misclassification error is $R(t)$, and its penalty cost of complexity is $\alpha \times 1$; Similarly, The subtree T_t means the subtree contains $|\tilde{T}_t|$ terminal nodes, and its misclassification error is $R(T_t)$, and its penalty cost of complexity is $\alpha \times |\tilde{T}_t|$.

In many cases, the misclassification error $R(t)$ is bigger than $R(T_t)$, the fact can be explained that the subtree T_t has more complex structure and then have better classification capability compared to subtree t . It also means that subtree T_t has better classification capability than subtree t .

However, if $R_\alpha(T_t) = R_\alpha(\{t\})$, the $\{t\}$ subtree is preferable because subtree t . and subtree T_t have the same sum of misclassification error and penalty cost of complexity. That is, although the subtree T_t has smaller misclassification error $R(T)$ than subtree t , after considering the penalty cost of complexity $\alpha \times |\tilde{T}_t|$, both of the two subtrees perform equivalently. According to the parsimonious rule, the subtree t is preferable.

In order to find the critical value , the following inequality is solved:

$$\begin{aligned} R_\alpha(T_t) &< R_\alpha(\{t\}) \\ \Rightarrow \alpha &< \frac{R(t) - R(T_t)}{|\tilde{T}_t|} \end{aligned}$$

Define function $g_t(t)$, where t belongs to T_t , as:

$$g_t(t) = \begin{cases} \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}, & t \notin \tilde{T}_t \\ +\infty, & t \in \tilde{T}_t \end{cases}$$

Define weakest link \bar{t}_1 in T_1 as:

$$g_1(\bar{t}_1) = \min_{t \in T_1} g_1(t)$$

$$\alpha_2 = g_1(\bar{t}_1)$$

The node \bar{t}_1 is the weakest link when the parameter α increases, \bar{t}_1 is also the first node such that $R(\{\bar{t}_1\})$ become equal to $R(T_1)$, and then the simple subtree $\{\bar{t}_1\}$ is preferable to the complex subtree T_1 , and α_2 is the value of α at which equality occurs.

Finally, a list of pruned $T(\alpha_k)$ trees can be obtained when α increases. The best pruned classification tree will be constructed.

2.4.2 Group Methods of Data Handling (GMDH)

Group Method of Data Handling (GMDH) [1] is applied in a great variety of areas in data mining. Inductive GMDH algorithms aim to find interrelations of variables in a data set and select the optimal structure of a model or a network. GMDH is an iterative method which successively tests models selected from a set of candidate models according to a specified criterion. General connection between input and output variables can be found in the form of a functional Volterra series, whose discrete analogue is known as the Kolmogorov-Gabor polynomial.

The polynomial can be expressed as follows,

$$y = a_0 + \sum_{i=1}^M a_i x_i + \sum_{i=1}^M \sum_{j=1}^M a_{ij} x_i x_j + \sum_{i=1}^M \sum_{j=1}^M \sum_{k=1}^M a_{ijk} x_i x_j x_k + \dots$$

where $X = (x_1, x_2, x_3, \dots, x_M)$ is the vector of input variables and

$A = (a_1, a_2, a_3, \dots, a_M)$ is the vector of summand coefficients.

The combinational GMDH algorithm has a multilayer iterative structure. Its specific feature is that the iteration rule does not remain consistent but expands with each new series. In the first series, all the models of the simplest structure are in the following form

$$y = a_0 + a_i x_i \quad i = 1, 2, \dots, M$$

After sorting these models, select the best F models by specified criterion. Models are sorted by series of equal structure complexity and best model is found for each series according to the specified criterion.

In the second series, models of more complex structure are sorted. These models are constructed on output variables from the best models of the first series:

$$y = a_0 + a_i x_i + a_2 x_j \quad i = 1, 2, \dots, F; \quad j = 1, 2, \dots, M. \quad F \leq M .$$

In the third series, the sorting involves more complex structure of the form as follows:

$$y = a_0 + a_i x_i + a_2 x_j + a_3 x_k$$

$$i = 1, 2, \dots, F; \quad j = 1, 2, \dots, F. \quad k = 1, 2, \dots, M \quad F \leq M .$$

The iterative procedure of the series continues until the criterion value stop increasing.

More complex iterative multilayered GMDH algorithm can be obtained by similar ways. The iteration rule remains the same for all series. For example, the form

$$y = a_0 + a_1 x_i + a_2 x_j + a_3 x_j x_i$$

is used in the first series, and a particular description

$$z = b_0 + b_1 y_i + b_2 y_j + b_3 y_j y_i$$

in the second series, and a particular description

$$w = c_0 + c_1 z_i + c_2 z_j + c_3 z_j z_i$$

is used in the third series, and so on. That is, the output values of a previous series are served as augments in the next series. The final model can be decided by specified external and internal criterion. The multilayered structure of GMDH algorithm can be shown in Fig.6.

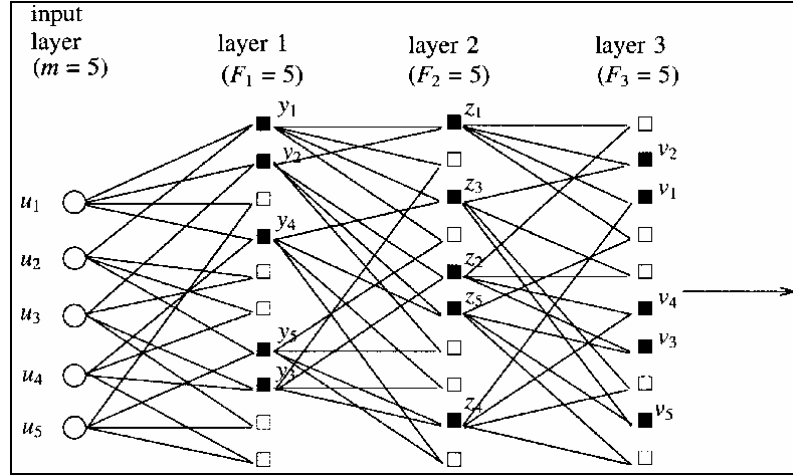


Figure 6 Multilayered structure of GMDH with five inputs and selected nodes

2.4.3 General Regression Neural Network (GRNN)

General regression neural network (GRNN) [7] is a one-pass learning network with a highly parallel structure. The algorithm can be used for any regression problems in which linearity assumption is not justified. GRNN provides estimates of continuous variables and converges to the underlying regression surface.

Suppose that $f(x, y)$ represents the known joint continuous probability density function. The regression of y on X is given by

$$E[y | X] = \frac{\int_{-\infty}^{\infty} y f(X, y) dy}{\int_{-\infty}^{\infty} f(X, y) dy}.$$

When the density $f(x, y)$ is unknown, it must be estimated from observations

of x and y . GRNN utilizes kernel density regression approach which adopted Parzen windows estimation

$$\hat{f}(X, Y) = \frac{1}{(2\pi)^{(p+1)/2} \sigma^{(p+1)}} \frac{1}{n} \sum_{i=1}^n \exp\left[-\frac{(X - X^i)^T (X - X^i)}{2\sigma^2}\right] \exp\left[-\frac{(Y - Y^i)^2}{2\sigma^2}\right]$$

to estimate $f(x, y)$. By using Parzen windows estimation, the GRNN estimator can

be easily presented as the following equation:

$$\hat{Y}(X) = \frac{\sum_{i=1}^n \exp\left[-\frac{(X - X^i)^T (X - X^i)}{2\sigma^2}\right] \int_{-\infty}^{\infty} y \exp\left[-\frac{(y - Y^i)^2}{2\sigma^2}\right] dy}{\sum_{i=1}^n \exp\left[-\frac{(X - X^i)^T (X - X^i)}{2\sigma^2}\right] \int_{-\infty}^{\infty} \exp\left[-\frac{(y - Y^i)^2}{2\sigma^2}\right] dy}$$

D_i^2 is defined as the a scalar function as follows:

$$D_i^2 = (X - X^i)^T (X - X^i)$$

Performing the substitution of D_i^2 yields the following GRNN estimator:

$$\hat{Y}(X) = \frac{\sum_{i=1}^n \exp\left[-\frac{D_i^2}{2\sigma^2}\right] Y^i}{\sum_{i=1}^n \exp\left[-\frac{D_i^2}{2\sigma^2}\right]}$$

Chapter 3 The Proposed Hybrid Model Approach

As mentioned in chapter 2, the feature selection problem is the main shortcoming in employing ANN to construct a neural network based credit scoring model. The hybrid model approach received a lot of attentions recently. Most of the studies on hybrid models are constructed by combining statistical method and ANN. The analytical procedure of credit scoring model proposed by this study mainly consists of two phases. In the first phase, the hybrid credit scoring model is composed of Classification and regression trees (CART) and other data mining algorithms such as BPN, LVQ, LDA etc. The first phase employs CART's predictive outcome and predictive categorical probability as input variables to construct the subsequent models using BPN, LDA, etc. The purpose of the first phase is to present a hybrid credit scoring model with higher accuracy and greater interpretable capability than the original credit scoring models without using hybrid approach. In the second phase, a predictive model of default period will be built through various data mining algorithms to obtain a precise estimator of default period. That is, for bad loaners, the time period between the loan start and the loan default is defined as the "default period". The objective of the second phase is to present a effective model to predict the default period of default-possible cases.

3.1 Model evaluation criterion

Financial loan companies often encounter considerable default loss due to misjudging or misclassifying the bad loan cases into "good loan" category. On the contrary, the loan companies will lose potential revenues if a good loan applicant is misclassified into "bad loan" category. The misclassified bad loan cases cause much greater loss to financial loan companies than misclassified good loan cases. Thus the prediction accuracy of "bad loan" is the higher the better for loan companies to

maintain an acceptable default risk. In this study, the good loan accuracy is specified to be greater than 50% to retain the essential profit.

3.2 Procedure of Constructing Hybrid Credit Scoring Model

Phase 1 : Construct Proposed Hybrid Credit Scoring Model

The proposed procedure of phase 1 can be shown in Fig. 7. Each step in phase 1 is described as follows:

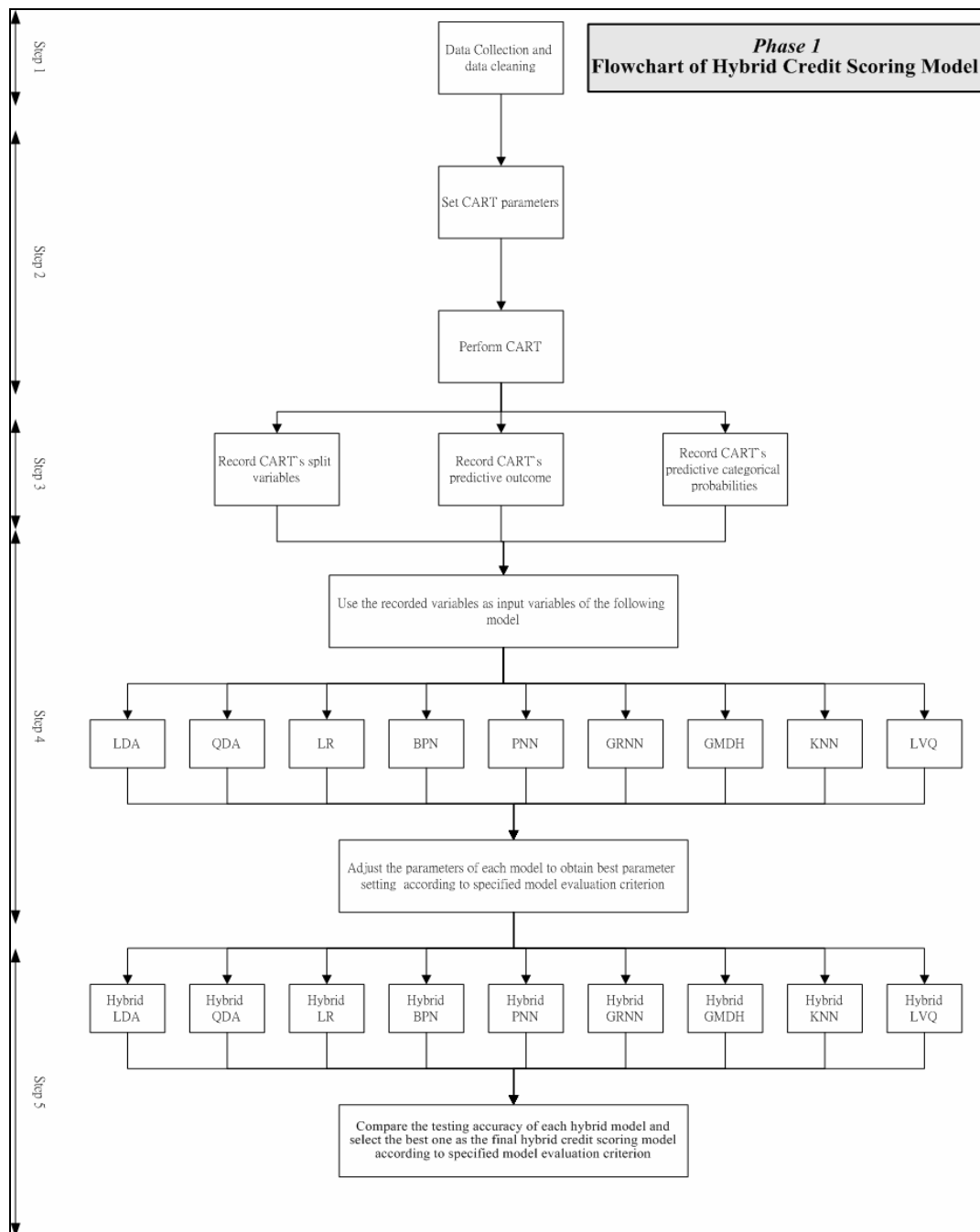


Figure 7. Flowchart of the proposed hybrid model

Step1: Data collection and cleaning

Loan applicants in this study are mid-sized companies whose financial statements are not as credible as those of public offering companies. Therefore, financial statement is only one part of considerable factors in this study. Loan companies usually adopt financial variables (quantitative factor) and non-financial variables (qualitative factor) simultaneously to increase model accuracy and reliability. This study collected loan data from a loan company in Taiwan in 2000 to 2003 as sample data and divided the dataset into two categories: “bad loan” and “good loan”. If a loan applicant is classified into “bad loan” category, the loan will be default and become a bad debt according to the proposed credit scoring model. On the contrary, “good loan” means the loan applicant can reimburse its debt in time.

Step2: Perform CART

The procedure of constructing CART can be described as follows:

Step 1. Decide impurity function.

Step 2. Grow tree by maximizing the decrease of tree impurity until the tree size becomes as large as possible.

Step 3. Prune tree structure.

Step 4. Use proper estimation method to obtain the honest estimator of tree classifier. The default setting is 10-fold cross validation.

Step 5. Interpret the results.

Step3: Record CART's split variables and predictive outcomes

In Step3, split variables of CART models can be deemed as the influential variables and should be recorded for further Steps. Similarly, CART's predictive

outcome and CART's predictive categorical probabilities can be deemed as important compressed information derived from CART model and should be retained as well. As a result, even the number of input variables of the hybrid model decreases, the model accuracy can still be retained using CART's predictive outcomes and CART's predictive categorical probabilities as input variables.

Step4: Use recorded variables and predictive outcomes as input variables of following model

CART has selected significant variables in Step3, therefore most of the relevant information are retained in the following three variables: "CART's predictive categorical probability of bad loan", "CART's predictive categorical probability of good loan" and "CART's predictive outcome". These variables can be used as augmented input variables of the subsequent model to enhance the accuracy of the hybrid model. Fig. 8 displays an example of CART's recorded variables which can be used as input variables of following BPN model. Similarly, these three recorded variables can be introduced to other algorithms such as LDA, LR, etc. This study also adopted many data mining algorithms to replace BPN to examine the effectiveness of proposed hybrid model. The cases given below described the credit scoring models constructed using the algorithm specified in each case.

Case 1. Linear Discriminant Analysis (LDA):

Specify appropriate prior probabilities for each category and utilize LDA to obtain results. LDA is performed using SAS 8.1 and the classification result is evaluated through N-fold cross validation.

Case 2. Quadratic Discriminant Analysis (QDA):

Specify appropriate prior probabilities for each category and utilize QDA to obtain results. QDA is performed using SAS 8.1 and the classification

result is evaluated through N-fold cross validation.

Case 3. Logistic Regression (LR):

Specify appropriate probability threshold value and utilize LR to obtain results. LR is performed using SAS 8.1 and the classification result is evaluated through hold-out method. 80% of data are chosen randomly to construct the LR model and the rest 20% of data are taken to validate the accuracy of LR model.

Case 4. Back Propagation Neural network (BPN):

The architecture of BPN [10] is decided to be three-layer BPN with completely interconnected neurons. With regard to the number of hidden nodes, this study adopted cascade learning rule to decide the proper number of hidden nodes. That is, cascade learning rule implies that hidden nodes increase gradually until the prediction accuracy of “testing bad loan” is not increased. As regards to the learning rate, momentum, and learning epochs, this study decided to use a small learning speed and long learning epochs to avoid the disturbance of overfitting. However, testing accuracy is another critical perspective when setting the number of epochs. The detail setting of network parameters are adhere to above principles. BPN is performed using Neural Shell2 (NeuralWare) and the classification result is evaluated through hold-out method. 80% of data are chosen randomly to train the BPN model and the rest 20% of data are used to validate the accuracy of BPN model.

Case 5. Probabilistic Neural network (PNN):

The architecture of PNN [10] can be easily determined from the observations of dataset. The only parameter which necessitates to be manually set is the smoothing parameter. This study adopts cascade

learning to decide best smoothing parameter. PNN is performed using Neural Shell2 (NeuralWare) and the classification result is evaluated through hold-out method. 80% of data are chosen randomly to train the PNN model and the rest 20% of data are used to validate the accuracy of PNN model.

Case 6. General Regression Neural network (GRNN):

The architecture of GRNN can also be easily determined from the observations of dataset as the same as PNN. The only parameter which necessitates manually setting is the smoothing parameter. This study here also adopts cascade learning to decide best smoothing parameter. GRNN is performed using Neural Shell2 (NeuralWare) and the classification result is evaluated through hold-out method. 80% of data are chosen randomly to train the GRNN model and the rest 20% of data are used to validate the accuracy of GRNN model.

Case 7. Group Method of Data Handling (GMDH):

GMDH is performed using Neural Shell2 (NeuralWare) and the classification result is evaluated through hold-out method. 80% of data are chosen randomly to train the GMDH model and the rest 20% of data are used to validate the accuracy of GMDH model.

Case 8. K-Nearest Neighbor (KNN):

It needs to set two parameters in training KNN [10], the first is the number of “K”, which represents the number of nearest neighbors, and the other is the measure of distance. This study utilizes Euclidean distance as measure of distance while performing KNN. As for the number “K”, rule of thumb (trial and error) method is employed to decide the best value for K. KNN is performed using Matlab6.5

(MathWorks inc) and the classification result is evaluated through hold-out method. 80% of data are chosen randomly to train the KNN model and the rest 20% of data are used to validate the accuracy of KNN model.

Case 9. Learning Vector Quantization (LVQ):

It needs to set three parameters mainly in training LVQ [10]. The first parameter is the number of prototypes, and another is learning rate and the other is the measure of distance. As for the number of initial prototypes, rule of thumb (trial and error) method is employed to decide the best value for the number of prototypes. Besides, the initial prototypes can be determined through random selection from the training samples. With respect to learning rate, preliminary experiments indicated the learning rate has no significant impacts for LVQ results. Hence this study set the value 0.1 as the learning rate. Similarly, this study utilizes Euclidean distance as measure of distance while performing LVQ. With respect to learning epochs, the number of learning epochs is not the critical factor in training LVQ because LVQ converges very fast. Thus the value of learning epochs is set to be 15. LVQ is performed using Matlab6.5 (MathWorks) and the classification result is evaluated through hold-out method. 80% of data are chosen randomly to train the LVQ model and the rest 20% of data are used to validate the accuracy of LVQ model.

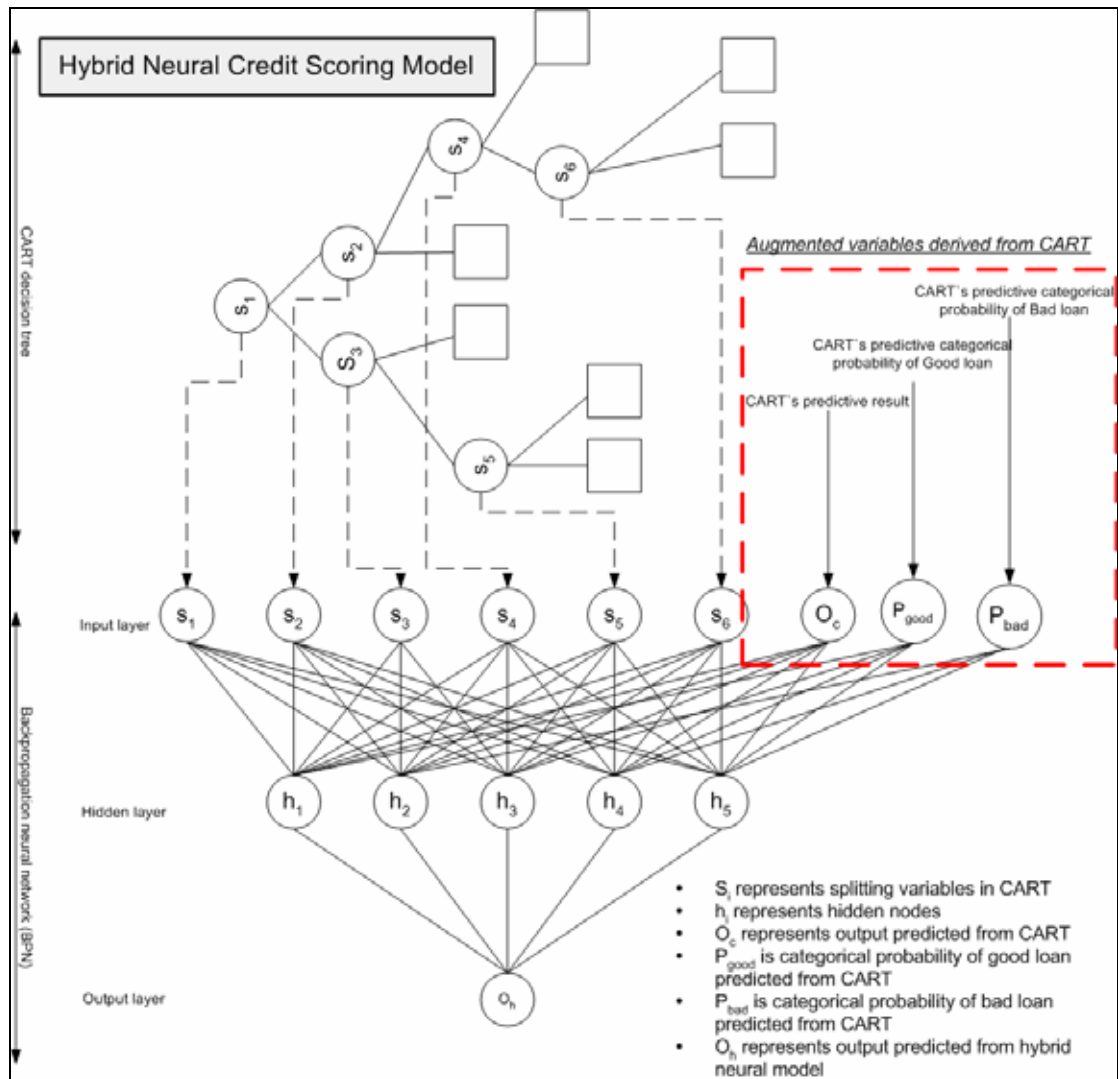


Figure 8. Hybrid BPN credit scoring model

Step5: Compare the accuracy of hybrid credit scoring model and select the best one as the final model.

The final credit scoring model is selected from the nine cases described in Step4. In other words, nine hybrid credit scoring models are constructed in Step4. According to the model evaluation criterion, select the best one as the final hybrid credit scoring model from the nine hybrid credit scoring models.

3.3 Establish Prediction Model of Default Period

Phase 2 : Establish Prediction Model of Default Period

For bad loaners, the time period between the loan start and the loan default is defined as the “default period”. Default period means the time period in which loaner still reimburse his debt regularly, the longer default period means the less potential profit loss to loan companies. On the contrary, the shorter default period represents the greater default risk. This phenomenon often makes loan companies unable to take proper reactions in time to the loan applicants with short default period.

Therefore, loan companies can take precautions and adopt corresponding reactions to the possible-default cases by reexamining the predicted default period when the loan applicant is classified into “Bad loaner category” in phase 1. Fig.9 describes the proposed procedure of phase 2.

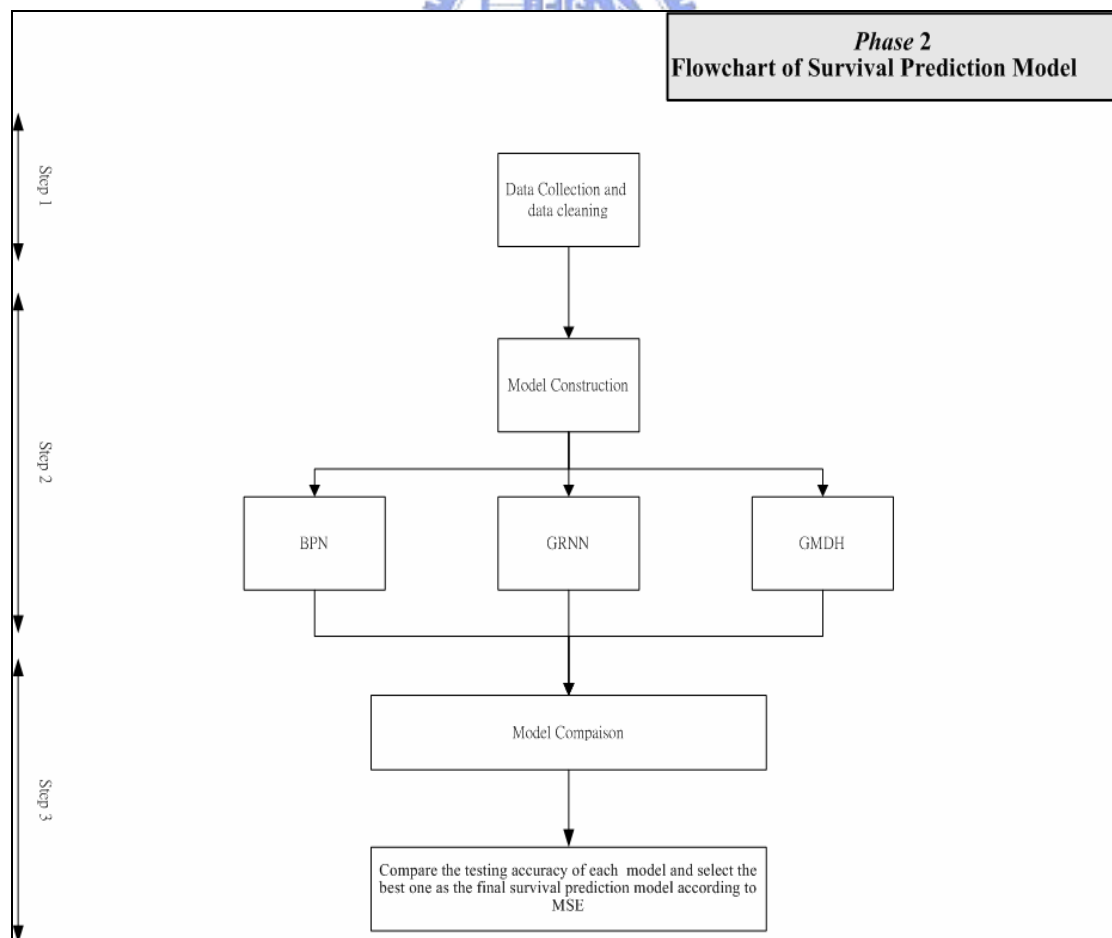


Figure.9 Flowchart of default period prediction model

Each step in phase2 is described as follows:

Step 1: Data collection and cleaning

The term “Default period” is only defined for bad loaners. The phase 2 simply choose bad loan data as sample data. Therefore, a prediction model of default period can be established through the bad loan cases. In addition, casewise deletion is adopted in this step.

Step 2: Model construction

This study employs three data mining algorithms to predict default period and the result of the three models are compared with the linear regression model. Three data mining algorithms are given below.

Case 1.Back Propagation Neural network (BPN):

The setting of parameters and network architecture are determined as mentioned in phase 1.

Case 2.General Regression Neural network (GRNN):

The setting of parameters and the GRNN network architecture are determined as mentioned in phase 1.

Case 3.Group Method of Data Handling (GMDH):

The setting of parameters and GMDH architecture are determined as mentioned in phase 1.

Step 3: Model comparison

The criterion for model comparison is mean square error (MSE). MSE is the smaller the better. The small MSE represents small difference between predicted output and the target. As a result, select the model with minimum value of testing MSE as the final model of default period. The MSE of linear regression is treated as a

benchmarking method in this step.



Chapter 4 Illustrative Examples

4.1 Description of Sample Data

The illustrative examples in this study consisted of 2080 commercial bank loaners, of which 1709 good loan cases and 371 bad loan cases. These data were obtained from a famous financial loan company in Taiwan for the period 2000 to 2003.

Each loan case included 31 variables of interest and some of these variables are non-financial variables. The variables are predetermined by the financial loan company. Detail descriptions of variables in the study are summarized in Table 3. It is noticeable that there are 14 financial variables and 17 non-financial variables, in which financial variables were directly measured from the financial statements and non-financial variables were indirectly measured by analysts' subjective determination. From the practical point of view, both financial and non-financial variables were used to construct the credit scoring model in this study.

Table 3. Variable Description

Variable Code	Rating Items	Variable Code	Rating Items
Financial structure (N6)	K83	Own capital rate	N1
	K85	Debit ratio	N2
	K87	Fix ratio	N3a
Liquidity Capability (N7)	K93	Current ratio	N3b
	K95	Rapid ratio	N4
	K97	DSR	N5
Management Capability	K100	Turnover days of Net value	N11
			Legal Policy

(N8)	K102	Turnover days of account receivable	N12	Economic Factor
	K104	Inventory Turnover	N13	Industry Trend
Profitability (N9)	K107	Gross profit rate	N14	Production capability
	K109	Net profit rate	N15	Marketing & Sales capability
	K111	EPS	N16	Management Teams capability
Growth Power (N10)	K114	Growth rate of EPS	N17	Evaluation by competitors and customers
	K116	Growth rate of sales volume	Net_Value	Net Value
	Default Period	Default Period	SCORE	Subtotal scores
	Capital	Capital of company		

4.2 Perform CART

This study used CART 5.0 sponsored by Salford systems to perform CART. After setting the minimum complexity α equal to zero and favor even split equal to 1, many preliminary experiments indicated that appropriate CART models can be obtained by adjusting prior probabilities shown in table 2.

Besides, this study repeats the proposed procedure of phase 1 six times to generate six different CART candidate models, and then use the six CART candidate models to construct the hybrid models. This practice intended to verify the effectiveness of the hybrid models produced by different CART candidate models. In

other words, if the hybrid model performs well under whichever the CART candidate models is selected, the hybrid model approach can be deemed as an effective methodology. Table 4 and displays the six different CART candidate models. The detail model of six CART candidate models can be found in Appendix.

Table 4. CART candidate models

CART model	Impurity function	Number of split variables	Testing Accuracy		Abbreviation of the model
			Good loan	Bad loan	
<i>Candidate1</i>	GINI	12	57.109	71.429	<i>Cart_1</i>
<i>Candidate2</i>	GINI	14	54.535	72.507	<i>Cart_2</i>
<i>Candidate3</i>	GINI	11	50.673	73.315	<i>Cart_3</i>
<i>Candidate4</i>	GINI	10	51.668	73.046	<i>Cart_4</i>
<i>Candidate5</i>	GINI	15	55.12	72.237	<i>Cart_5</i>
<i>Candidate6</i>	GINI	9	51.551	73.315	<i>Cart_6</i>

The split variables of each produced CART model are listed in table 5. Significant reduction of input variables can be observed in table 5. Furthermore, these split variables can be regarded as influential variables and be used to construct the hybrid model.

Table 5. CART's split variables

CART model	Number of split variables	Split variables
<i>Candidate1</i>	12	<i>N4 N5 N6 N7 N14 K95 K97 K104 K107 K109 Capital Net_Value</i>
<i>Candidate2</i>	14	<i>N3 N4 N5 N6 N7 N9 N14 N15 K104 K107 K109 K116 Capital Net_Value</i>
<i>Candidate3</i>	11	<i>N4 N6 N9 N14 N15 N17 K85 K97 K109 Capital Net_Value</i>

<i>Candidate4</i>	10	<i>N4 N6 N9 N14 N15 K85 K97 K109 Capital Net_Value</i>
<i>Candidate5</i>	15	<i>N3 N4 N6 N7 N9 N14 N15 K87 K95 K97 K104 K107 K109 K116 Net_Value</i>
<i>Candidate6</i>	9	<i>N4 N6 N7 N14 N15 N17 K85 K97 Net_Value</i>

Apparently, the original CART does not provide satisfactory results under anyone of the six candidate models.

Other original credit scoring models were also established and summarized in Table 6 as benchmarking methods. This study adopted an extensive trial and error method to find the best parameter setting for each model. After many preliminary experiments, the best parameter setting and testing accuracy of each original model can be obtained and showed in Table 6.

Table 6. Comparison of original credit scoring models

Model	Abbreviation	Testing Accuracy (%)		Notes
		Bad Loan	Good Loan	
Linear Discriminant Analysis	<i>LDA</i>	79.51	50.46	Priors: 0.63 :0.37
Quadratic Discriminant Analysis	<i>QDA</i>	76.01	51.61	Priors: 0.66 :0.34
Logistic Regression	<i>LR</i>	79.2	51	Probability level: 0.12
Classification & Regression Tree	<i>CART</i>	73.04	51.66	Priors: 0.59 :0.41
Probabilistic Neural Network	<i>PNN</i>	52.05	77.26	Smoothing factor 0.355
Backpropagation Neural Network	<i>BPN</i>	82.19	51.31	Hidden node: 15
General Regression Neural Network	<i>GRNN</i>	81.03	62.56	Smoothing factor 0.6583
Group Method of Data Handling	<i>GMDH</i>	82.27	50.74	Criterion value 0.150836
K-Nearest Neighbor	<i>KNN</i>	25.28	86.93	K=1

Learning Vector Quantization	<i>LVQ</i>	29.88	93.27	Prototypes: 400
------------------------------	------------	-------	-------	-----------------

4.3 Record CART's Split Variables and Predictive Outcomes

Spilt variables, predictive categorical probabilities and predictive result of CART were recorded and used as input variables for the further hybrid models.

4.4 Use recorded variables and predictive outcomes as input variables of following model

The three variables: spilt variables, predictive categorical probabilities and predictive result were used as input variables in LDA, QDA, BPN, etc. The input variables of the following hybrid model are summarized in Table 7.

Table 7. Input Variables of following hybrid models

CART model	Input variables of following hybrid models	
	Split variables	Augmented variables from CART model
<i>Candidate1</i>	<i>N4 N5 N6 N7 N14 K95 K97 K104 K107 K109 Capital Net_Value</i>	<i>Predictive probability of bad loan of CART.</i>
<i>Candidate2</i>	<i>N3 N4 N5 N6 N7 N9 N14 N15 K104 K107 K109 K116 Capital Net_Value</i>	
<i>Candidate3</i>	<i>N4 N6 N9 N14 N15 N17 K85 K97 K109 Capital Net_Value</i>	<i>Predictive probability of good loan of CART.</i>
<i>Candidate4</i>	<i>N4 N6 N9 N14 N15 K85 K97 K109 Capital Net_Value</i>	<i>Predictive outcome of CART.</i>
<i>Candidate5</i>	<i>N3 N4 N6 N7 N9 N14 N15 K87 K95 K97 K104 K107 K109 K116 Net_Value</i>	
<i>Candidate6</i>	<i>N4 N6 N7 N14 N15 N17 K85 K97 Net_Value</i>	

The procedure of constructing various hybrid models followed the principles described in Step 4 in section 3.2. This study used SAS 8.1 to perform LDA, QDA

and LR analysis. According to each CART candidate model, a corresponding hybrid model was built and shown as table 8.

Case 1. Linear Discriminant Analysis (LDA):

The performance of hybrid LDA model and the original LDA model was compared and the results were listed in Table 8 and Table 9. Obviously, the accuracy of hybrid LDA model for the testing bad loan was significantly higher than the original LDA by 5% no matter which CART candidate model was used.

Table 8. Hybrid LDA Performance

Hybrid LDA			
Hybrid LDA model	Prior Bad:Good	N-fold CV accuracy(%)	
		Bad loan	Good loan
Cart_1	0.69:0.31	83.02	50.2
Cart_2	0.61:0.39	88.14	54.18
Cart_3	0.70:0.3	85.41	51.96
Cart_4	0.66:0.34	84.59	50.19
Cart_5	0.64:0.36	82.43	51.96
Cart_6	0.66:0.34	82.7	51.08

Table 9. Original LDA Performance

Original LDA			
LDA model	Priors Bad:Good	N-fold CV accuracy(%)	
		Bad loan	Good loan
LDA-1	0.60:0.40	77.9	55.12
LDA-2	0.61:0.39	78.44	53.66
LDA-3	0.62:0.38	78.71	52.31
LDA-4	0.63:0.37	79.51	50.46
LDA-5	0.64:0.36	80.32	48.57
The best five LDA models			

Case 2. Quadratic Discriminant Analysis (QDA):

Table 10 and Table 11 also indicated that the hybrid QDA had better prediction accuracy than the original QDA model. Obviously, the testing bad loan accuracy of hybrid QDA model increased at least by 7% compared to the original QDA no matter which CART candidate model was used.

Table 10. Hybrid QDA Performance

Hybrid QDA

Table 11. Original QDA Performance

Original QDA

<i>Hybrid QDA model</i>	<i>Prior Bad:Good</i>	<i>N-fold CV accuracy(%)</i>		<i>QDA Model</i>	<i>Priors Bad:Good</i>	<i>N-fold CV accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>			<i>Bad loan</i>	<i>Good loan</i>
<i>Cart_1</i>	<i>0.73:0.27</i>	<i>83.02</i>	<i>59.74</i>	<i>QDA-1</i>	<i>0.62:0.38</i>	<i>73.85</i>	<i>57.99</i>
<i>Cart_2</i>	<i>0.74:0.26</i>	<i>83.29</i>	<i>59.57</i>	<i>QDA-2</i>	<i>0.63:0.37</i>	<i>74.66</i>	<i>56.52</i>
<i>Cart_3</i>	<i>0.50:0.5</i>	<i>82.16</i>	<i>49.39</i>	<i>QDA-3</i>	<i>0.64:0.36</i>	<i>75.2</i>	<i>55</i>
<i>Cart_4</i>	<i>0.52:0.48</i>	<i>82.16</i>	<i>50.85</i>	<i>QDA-4</i>	<i>0.65:0.35</i>	<i>75.74</i>	<i>53.13</i>
<i>Cart_5</i>	<i>0.50:0.5</i>	<i>82.97</i>	<i>45.64</i>	<i>QDA-5</i>	<i>0.66:0.34</i>	<i>76.01</i>	<i>51.61</i>
<i>Cart_6</i>	<i>0.56:0.44</i>	<i>84.05</i>	<i>50.53</i>	<i>The best five QDA models</i>			

Case 3. Logistic Regression (LR):

Similar results as in Case 1 and Case 2 can be observed in Table 12 and Table 13. This also indicated that the hybrid LR model significantly performed better than the original LR model according to the specified model evaluation criterion.

Table 12. Hybrid LR Performance

Table 13. Original LR Performance

<i>Hybrid LR</i>			
<i>Hybrid LR model</i>	<i>Probability level</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>
<i>Cart_1</i>	<i>0.12</i>	<i>84.1</i>	<i>50</i>
<i>Cart_2</i>	<i>0.08</i>	<i>87.1</i>	<i>57.6</i>
<i>Cart_3</i>	<i>0.10</i>	<i>83.2</i>	<i>57.6</i>
<i>Cart_4</i>	<i>0.12</i>	<i>83.2</i>	<i>55.4</i>
<i>Cart_5</i>	<i>0.12</i>	<i>81.4</i>	<i>55.5</i>
<i>Cart_6</i>	<i>0.08</i>	<i>88.1</i>	<i>51</i>

<i>Original LR</i>			
<i>LR model</i>	<i>Probability level</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>
<i>LR-1</i>	<i>0.12</i>	<i>79.2</i>	<i>51</i>
<i>LR-2</i>	<i>0.14</i>	<i>75.5</i>	<i>57.8</i>
<i>LR-3</i>	<i>0.16</i>	<i>72.2</i>	<i>65.2</i>
<i>LR-4</i>	<i>0.18</i>	<i>68.2</i>	<i>70.9</i>
<i>LR-5</i>	<i>0.20</i>	<i>65.2</i>	<i>75.6</i>
<i>The best five LR models</i>			

Case 4. Back Propagation Neural network (BPN):

The procedure of BPN can be stated as follows: a very small learning rate at 0.001, momentum as 0.85, and learning epoch as 2000 are set in the BPN training period to avoid overfitting problem and fluctuation of predictive accuracy. With regard to the number of hidden nodes, this study adopted cascade learning rule to decide the proper number of hidden nodes. Cascade learning rule implies that hidden nodes increase gradually until the accuracy of testing bad loan stop increasing. For instance, the results of cascade learning procedure were plotted in Fig.10 and Fig.11. Moreover, Fig.10 and Fig.11 also indicated that the prediction accuracy of hybrid BPN model produced by Cart_1 increased up to 10% as compared to the original BPN model. Other hybrid BPN models also have the same improvement on the bad loan accuracy.

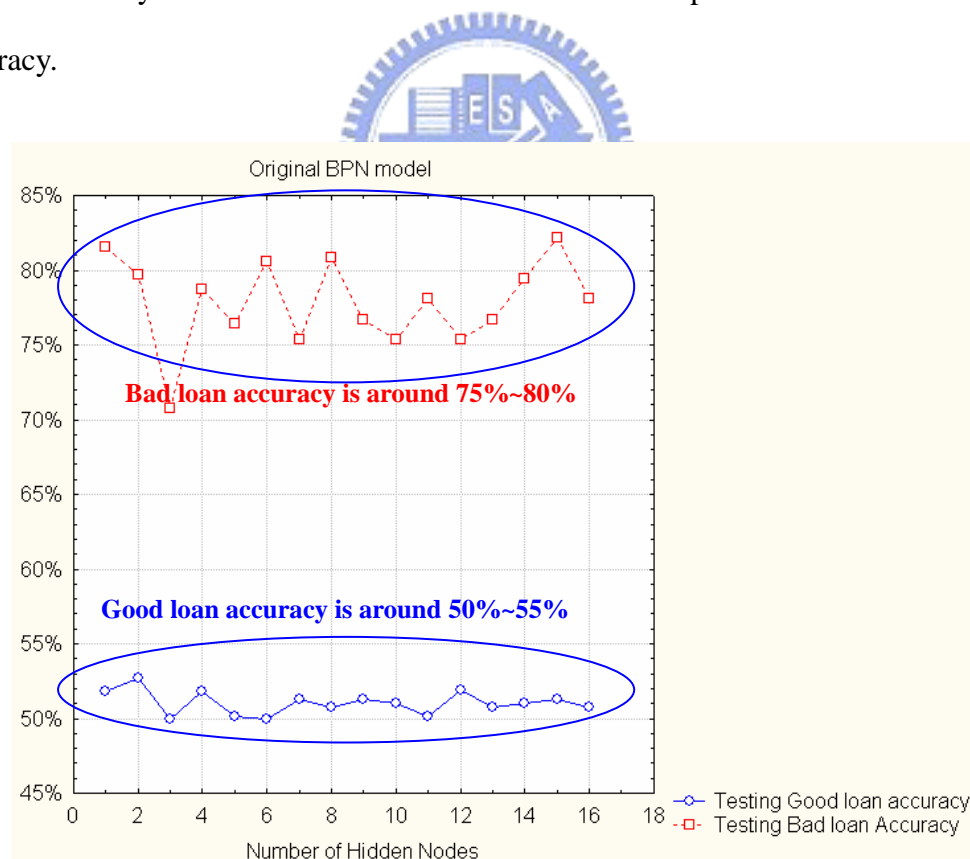


Fig.10. Original BPN model accuracy

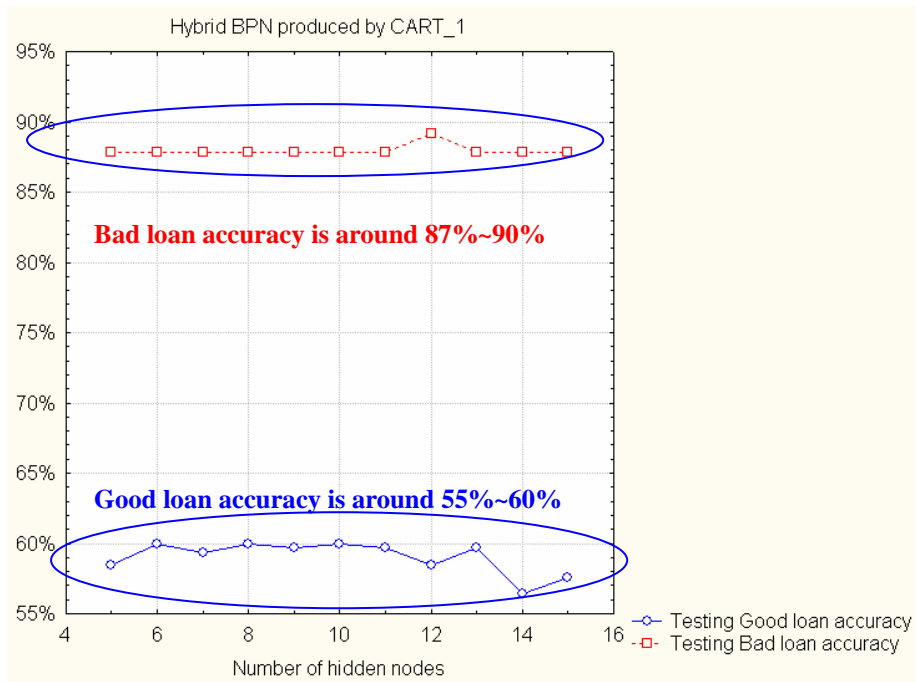


Fig.11. Hybrid BPN produced by Cart_1

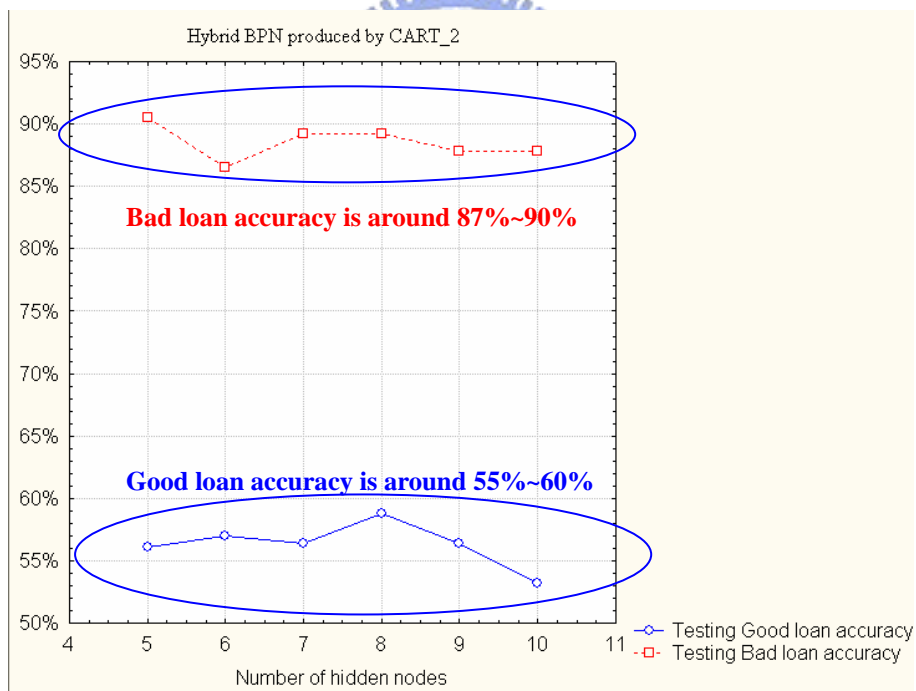


Fig.12. Hybrid BPN produced by Cart_2

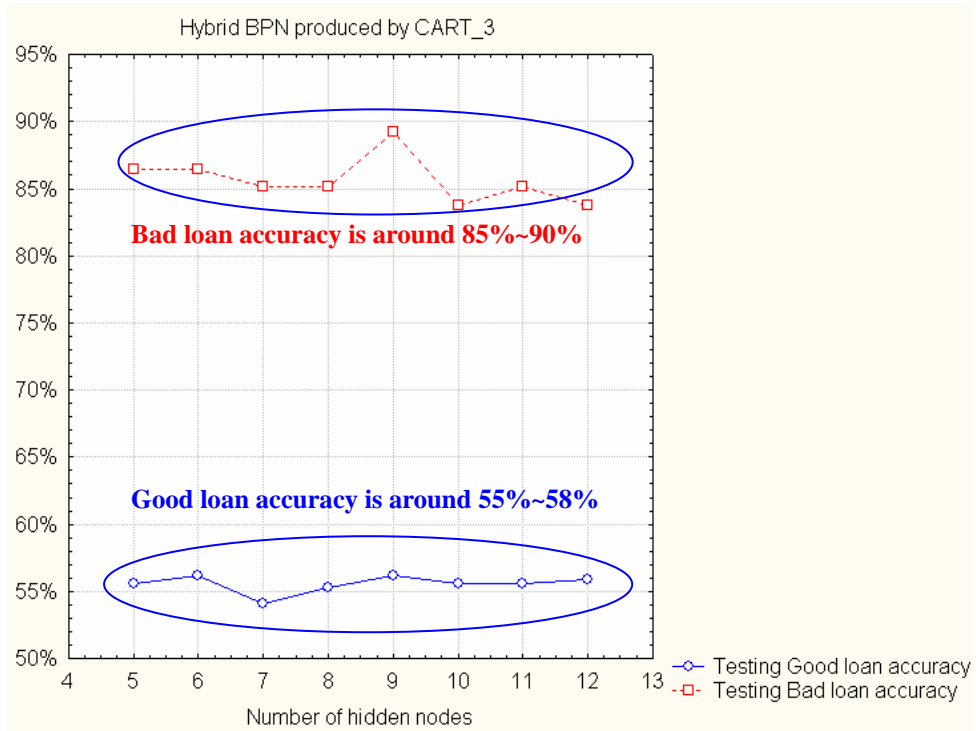


Fig.13. Hybrid BPN produced by Cart_3

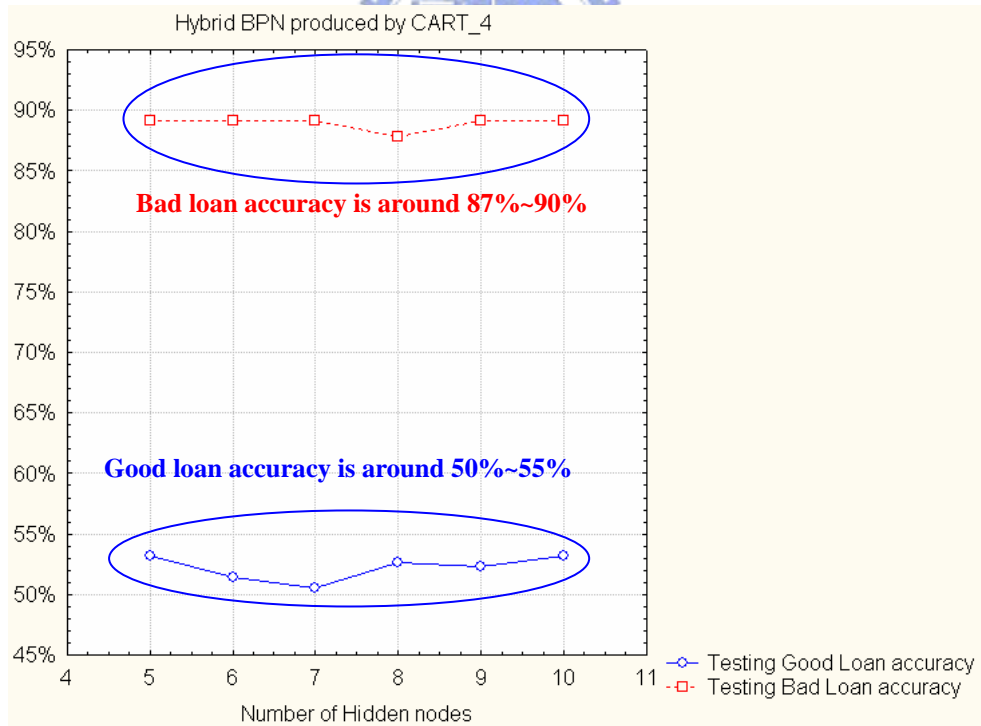


Fig.14. Hybrid BPN produced by Cart_4

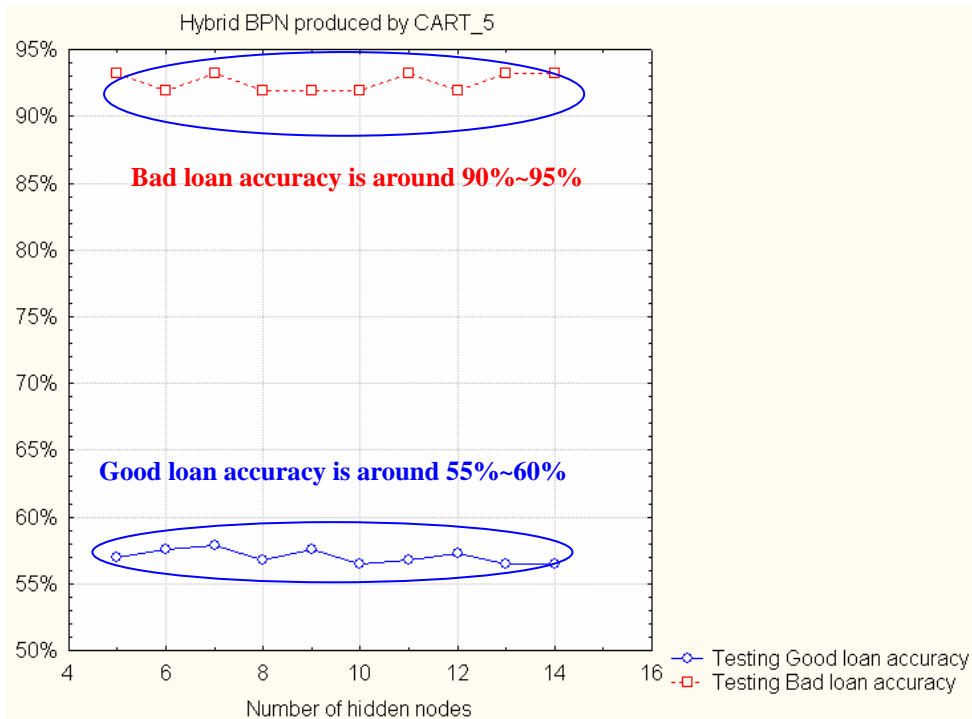


Fig.15. Hybrid BPN produced by Cart_5

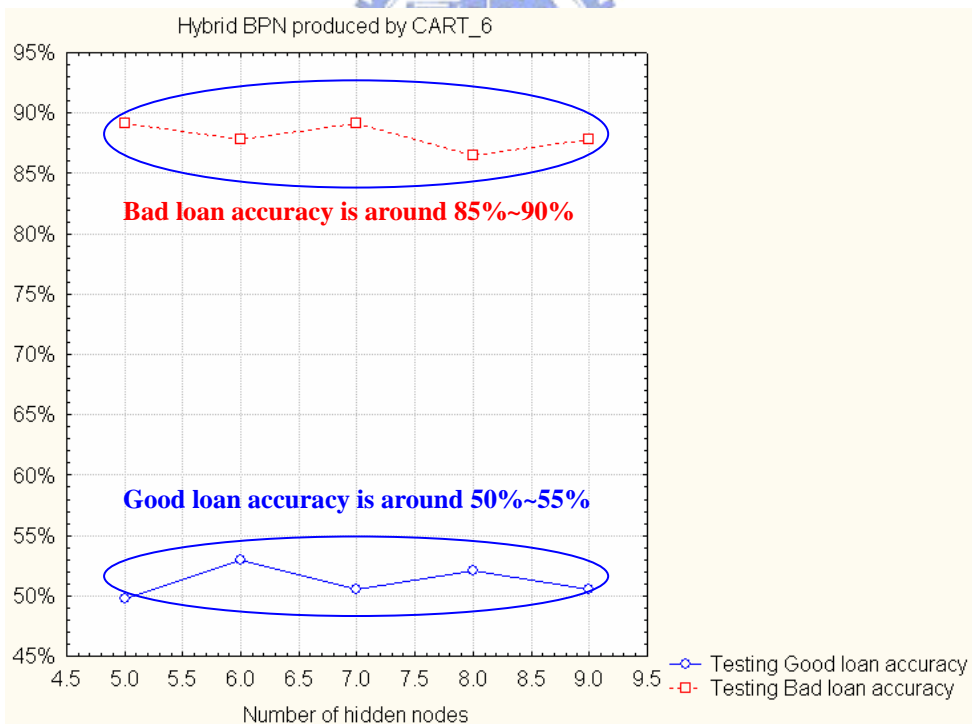


Fig.16. Hybrid BPN produced by Cart_6

Obviously, Fig.11-Fig.16 indicated the significant effectiveness by using hybrid model approach. The prediction accuracy of bad loan increases at least by 10%~15%.

The accuracy of good loan also increases by 5%. The performance of the proposed hybrid BPN exceeds what we expected according to specified model evaluation criterion.

Case 5. Probabilistic Neural network (PNN):

Even Probabilistic Neural network (PNN) is adopted, the same improvement of prediction accuracy can be obtained in Table 14 and Table 15. The results in these tables also indicated that hybrid PNN model performed significantly better than the original PNN model.

Table 14. Hybrid PNN Performance

<i>Hybrid PNN</i>			
<i>Hybrid PNN model</i>	<i>Smoothing factor</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>
<i>Cart_1</i>	<i>0.1808</i>	<i>71.23</i>	<i>81.87</i>
<i>Cart_2</i>	<i>0.1730</i>	<i>63.01</i>	<i>90.53</i>
<i>Cart_3</i>	<i>0.1500</i>	<i>56.06</i>	<i>76.44</i>
<i>Cart_4</i>	<i>0.2545</i>	<i>72.97</i>	<i>86.26</i>
<i>Cart_5</i>	<i>0.1691</i>	<i>64.86</i>	<i>84.41</i>
<i>Cart_6</i>	<i>0.2118</i>	<i>63.51</i>	<i>89.71</i>

Table 15. Original PNN Performance

<i>Original PNN</i>			
<i>PNN model</i>	<i>Smoothing factor</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>
<i>PNN-1</i>	<i>0.2375</i>	<i>33.33</i>	<i>85.88</i>
<i>PNN-2</i>	<i>0.2515</i>	<i>42.86</i>	<i>80.92</i>
<i>PNN-3</i>	<i>0.2375</i>	<i>34.38</i>	<i>82.62</i>
<i>PNN-4</i>	<i>0.2375</i>	<i>31.67</i>	<i>84.46</i>
<i>PNN-5</i>	<i>0.3550</i>	<i>52.05</i>	<i>77.26</i>
<i>The best five PNN models</i>			

Case 6. General Regression Neural network (GRNN):

As compared to the original credit scoring models, GRNN performed best among all original models. The performance of hybrid GRNN model is still quite good. Almost 5% to 10% accuracy improvement was obtained when hybrid GRNN model was employed. Table 16 and Table 17 indicated that hybrid GRNN model performed significantly better than the original models.

Table 16. Hybrid GRNN Performance

<i>Hybrid GRNN</i>			
<i>Hybrid GRNN model</i>	<i>Smoothing factor</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good Loan</i>
<i>Cart_1</i>	<i>0.2972</i>	<i>87.83</i>	<i>56.72</i>
<i>Cart_2</i>	<i>0.4370</i>	<i>93.24</i>	<i>57.89</i>
<i>Cart_3</i>	<i>0.3205</i>	<i>90.54</i>	<i>56.14</i>
<i>Cart_4</i>	<i>0.4836</i>	<i>86.48</i>	<i>63.15</i>
<i>Cart_5</i>	<i>0.3128</i>	<i>82.18</i>	<i>61.26</i>
<i>Cart_6</i>	<i>0.3943</i>	<i>81.08</i>	<i>64.91</i>

Table 17. Original GRNN Performance

<i>Original GRNN</i>			
<i>GRNN model</i>	<i>Smoothing factor</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>
<i>GRNN-1</i>	<i>0.6777</i>	<i>81.03</i>	<i>57.82</i>
<i>GRNN-2</i>	<i>0.6661</i>	<i>79.31</i>	<i>62.01</i>
<i>GRNN-3</i>	<i>0.6816</i>	<i>81.03</i>	<i>59.77</i>
<i>GRNN-4</i>	<i>0.6816</i>	<i>81.03</i>	<i>60.33</i>
<i>GRNN-5</i>	<i>0.6505</i>	<i>81.03</i>	<i>62.29</i>
<i>The best five GRNN models</i>			

Case 7. Group Method of Data Handling (GMDH):

GMDH did not perform well as compared to the original models. However, hybrid GMDH model had surprisingly promising accuracy in all GMDH hybrid model. Almost 10% accuracy improvement was obtained when hybrid GMDH was employed. Table 18 and Table 19 indicated that the hybrid GMDH model performed significantly better than original models.

Table 18. Hybrid GMDH performance

<i>Hybrid GMDH</i>		
<i>Hybrid GMDH model</i>	<i>Testing accuracy(%)</i>	
	<i>Bad loan</i>	<i>Good loan</i>
<i>Cart_1</i>	<i>87.83</i>	<i>50.87</i>

Table 19. Original GMDH performance

<i>Original GMDH</i>		
<i>GMDH model</i>	<i>Testing accuracy(%)</i>	
	<i>Bad loan</i>	<i>Good loan</i>
<i>GMDH-1</i>	<i>74.13</i>	<i>49.72</i>

<i>Cart_2</i>	<i>91.89</i>	<i>56.43</i>		<i>GMDH-2</i>	<i>82.27</i>	<i>50.74</i>
<i>Cart_3</i>	<i>87.83</i>	<i>57.01</i>		<i>GMDH-3</i>	<i>80.95</i>	<i>51.27</i>
<i>Cart_4</i>	<i>90.54</i>	<i>53.50</i>		<i>GMDH-4</i>	<i>77.77</i>	<i>50.99</i>
<i>Cart_5</i>	<i>89.18</i>	<i>55.26</i>		<i>GMDH-5</i>	<i>73.01</i>	<i>52.12</i>
<i>Cart_6</i>	<i>90.54</i>	<i>51.46</i>		<i>The best five GMDH models</i>		

Case 8. K-Nearest Neighbor (KNN):

KNN did not result in satisfactory result. The prediction accuracy of bad loan for KNN models is far lower than 50% which is requested by the model evaluation criterion. Although the total accuracy of both good loan and bad loan is still quite promising, the fact that KNN can't incorporate different objectives of various categories make KNN hard to be applied in constructing the credit scoring model. Similarly, the performance of hybrid KNN models also produced disappointed results. Theoretically, the possible reason for the poor classification capability of KNN might be inferred to the extremely gap of sample sizes between bad loan class and good loan class. Prototype methods such as KNN classify observation according to the major class of "K" nearest neighbors. That is, if the difference of sample size between bad loan class and good loan class become extremely big, the "K" nearest neighbors might all belong to the same category. However the phenomenon is not induced by the general KNN classification rule but induced by the extremely difference of sample size between bad loan class and good loan class.

Table 20. Hybrid KNN Performance

<i>Hybrid KNN</i>			
<i>Hybrid KNN model</i>	<i>Neighbor number K</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>

Table 21. Original KNN Performance

<i>Original KNN</i>			
<i>KNN model</i>	<i>Neighbor number K</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>

<i>Cart_1</i>	<i>1</i>	<i>36.48</i>	<i>86.55</i>	<i>KNN-1</i>	<i>1</i>	<i>25.28</i>	<i>86.93</i>
<i>Cart_2</i>	<i>1</i>	<i>22.89</i>	<i>86.48</i>	<i>KNN-2</i>	<i>2</i>	<i>25.28</i>	<i>86.93</i>
<i>Cart_3</i>	<i>1</i>	<i>30.45</i>	<i>86.45</i>	<i>KNN-3</i>	<i>3</i>	<i>17.24</i>	<i>96.04</i>
<i>Cart_4</i>	<i>1</i>	<i>32.14</i>	<i>87.04</i>	<i>KNN-4</i>	<i>4</i>	<i>17.24</i>	<i>95.75</i>
<i>Cart_5</i>	<i>1</i>	<i>29.11</i>	<i>87.53</i>	<i>KNN-5</i>	<i>5</i>	<i>14.94</i>	<i>97.87</i>
<i>Cart_6</i>	<i>1</i>	<i>33.33</i>	<i>88.06</i>	<i>The best five KNN models</i>			

Case 9. Learning Vector Quantization (LVQ):

The performance of LVQ is as poor as that of KNN. The reason might be the same as for the poor performance of KNN. Both LVQ and KNN are prototype methods theoretically, as a result, the similar depressed result of LVQ is likely to be anticipated. Table 22 and Table 23 showed the performance of hybrid LVQ and original LVQ models. However, 10% to 15% improvement in prediction accuracy for bad loan can be still observed in Table 22.

Table 22. Hybrid LVQ Performance

<i>Hybrid LVQ</i>			
<i>Hybrid LVQ model</i>	<i>Prototype number K</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>
<i>Cart_1</i>	<i>250</i>	<i>27.02</i>	<i>94.44</i>
<i>Cart_2</i>	<i>150</i>	<i>31.08</i>	<i>93.86</i>
<i>Cart_3</i>	<i>200</i>	<i>37.87</i>	<i>92</i>
<i>Cart_4</i>	<i>200</i>	<i>33.33</i>	<i>88.82</i>
<i>Cart_5</i>	<i>250</i>	<i>27.84</i>	<i>87.24</i>

Table 23. Original LVQ Performance

<i>Original LVQ</i>			
<i>LVQ model</i>	<i>Prototype number K</i>	<i>Testing accuracy(%)</i>	
		<i>Bad loan</i>	<i>Good loan</i>
<i>LVQ-1</i>	<i>100</i>	<i>26.43</i>	<i>92.40</i>
<i>LVQ-2</i>	<i>250</i>	<i>27.58</i>	<i>91.18</i>
<i>LVQ-3</i>	<i>300</i>	<i>22.98</i>	<i>92.09</i>
<i>LVQ-4</i>	<i>350</i>	<i>22.98</i>	<i>91.18</i>
<i>LVQ-5</i>	<i>400</i>	<i>29.88</i>	<i>92.70</i>

<i>Cart_6</i>	<i>300</i>	<i>23.45</i>	<i>91.94</i>	<i>The best five LVQ models</i>
---------------	------------	--------------	--------------	---------------------------------

4.5 Compare the accuracy of each hybrid model and choose the best hybrid credit scoring model

The final hybrid credit scoring model can be easily derived from the Table 24.

Table 24. Best Hybrid Credit Scoring Model

<i>Hybrid Model</i>	<i>Method</i>	<i>Bad loan Accuracy</i>	<i>Good loan Accuracy</i>	<i>Note</i>	<i>Rank</i>
<i>Cart_2+GRNN</i>	<i>CART + GRNN</i>	<i>93.24</i>	<i>57.89</i>	<i>Smoothing factor: 0.4370</i>	<i>1st</i>
<i>Cart_2+ BPN</i>	<i>CART + BPN</i>	<i>93.24</i>	<i>57.01</i>	<i>Hidden nodes: 5</i>	<i>2nd</i>
<i>Cart_2+ GMDH</i>	<i>CART + GMDH</i>	<i>91.89</i>	<i>56.43</i>	<i>Criterion Value: 0.1396</i>	<i>3rd</i>

The Hybrid model “Cart_2+GRNN” is the final hybrid credit scoring model with the highest prediction accuracy. Nearly 15% improvement in prediction accuracy of bad loan was obtained when the original best model was compared with this hybrid model.

The result of proposed hybrid model demonstrated the fact that no matter which the CART candidate model is selected, or whatever the following algorithms is utilized, the prediction accuracy of the proposed hybrid model is always significantly higher than original models. The result also strongly support that CART can be used as the feature selection method to enhance the classification accuracy.

4.6 Establish Prediction Model of Default Period

The term “Default period” is only defined for bad loaners. This study chose bad

loan data as the sample data to construct the prediction model of default period. Therefore, the prediction model of default period can be established through the bad loan cases in phase 2. In addition, casewise deletion is adopted here. The testing MSE of each prediction model of default period is summarized in Table 25. The prediction model with the minimum testing MSE was selected as the final prediction model of default period. The procedure and principles of constructing each prediction model were the same as described in the previous Sections. The stepwise linear regression model is also constructed as a benchmarking method and the comparisons of various prediction models of default period are shown in Table 25.

Table 25. The MSE of each prediction model

<i>Prediction Model</i>	<i>MSE</i>	<i>Rank</i>
<i>GMDH</i>	<i>20.354</i>	<i>1st</i>
<i>BPN</i>	<i>21.927</i>	<i>2nd</i>
<i>GRNN</i>	<i>23.468</i>	<i>3rd</i>
<i>Linear Regression</i>	<i>32.678</i>	<i>4th</i>

According to Table 25, GMDH is chosen to be the final prediction model of default period. Hence, the default period can predict more precisely by GMDH than any other models.


4.6 Further Comparison of Hybrid Model

We are now in a position to say the fact that using CART as a preprocessing mechanism or feature selection tool can definitely increase the model accuracy. The results derived from many algorithms have verified the generalization capability of the proposed hybrid CART model. The concept of proposed model can be easily applied in other classification methods such as Support Vector Machine (SVM) [10, 11]. However, it is still a doubtful point: Can we use other algorithms rather than

CART? Shall the performance of other classification tools (Such as LDA) perform better than the proposed hybrid model built from CART?

As the question noted earlier, the study explored the hybrid model which adopted LDA as feature selection method in phase1. The similar procedure mentioned in chapter 3 is utilized to construct the hybrid model derived from LDA. First, the variables selected by LDA, the LDA's predicted outcome and LDA's predicted categorical probabilities were adopted as input variables of the following models. These chosen variables of LDA are listed in Table 26.

Table 26. Variables selected by LDA

LDA Model	Input Variables of following hybrid models	
	Discriminator Variables	Augmented Variables from LDA
LDA	 <i>N1 N2 N4 N5 N7 N13 N14 N15 K85 K87 K97</i> <i>K109 Net_Value</i>	<i>Posterior probability of bad loan of LDA.</i> <i>Posterior probability of good loan of LDA.</i> <i>Predictive outcome of LDA.</i>

In addition, Table 27 indicated the result of LDA based hybrid model has inferior accuracy than CART based hybrid model no matter what the subsequent model was used. 5% to 10% degradation of accuracy can be observed in Table 27.

Table 27. The performance of hybrid model based on LDA

Hybrid Model	Bad loan Accuracy	Good loan Accuracy	Compare to Original Model	Compare to Hybrid model based on CART
LDA+BPN	84.93	53.06	Better	Worse

<i>LDA+GRNN</i>	<i>83.56</i>	<i>50.72</i>	<i>Better</i>	<i>Worse</i>
<i>LDA+GMDH</i>	<i>81.08</i>	<i>52.92</i>	<i>Better</i>	<i>Worse</i>

Consider the illustrative example given above, the hybrid model based on CART does have better performance than LDA based models. Although the prediction accuracy of hybrid model based on LDA increases slightly more than the original models, this study still recommend to use the proposed model to obtain accurate prediction results.



Chapter 5 Concluding Remarks

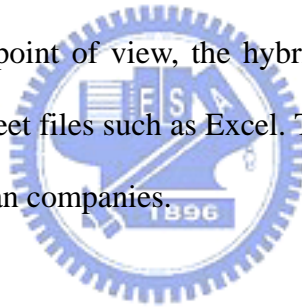
For loan companies, establishing a reliable credit scoring model can significantly reduce their default risk and increase the profits, especially in the economic recession environment. Many banks and loan companies devoted themselves to developing internal credit scoring model according to the New Basel Capital Accord and the logistic regression credit scoring model is usually the first priority choice for banks and loan companies in Taiwan.

This study presents a new hybrid approach to obtain superior classification accuracy and interpretation capability than the conventional credit scoring models. The feasibility of the proposed hybrid credit scoring was demonstrated by an illustrative example and the effectiveness of the proposed hybrid approach was verified through an extensive comparison with various hybrid models produced from six different CART trees. The proposed hybrid credit scoring model possesses good interpretable capability through identifying CART's split variables. Using the variables chosen by CART would include critical information of loan applicants and the decision makers of loan companies can make correct judgment based upon the proposed model.

In addition, this study presented an extensive comparison among various data mining algorithms applied in constructing the credit scoring models. This study also made comparisons among various hybrid credit scoring models. By adopting the proposed hybrid model, loan companies can establish their own reliable credit scoring model with high accuracy and good interpretable capability. The proposed hybrid model can reduce possible default risks and increase considerable amount of profits.

The contribution of this study can be summarized as follows:

1. The study provides a new hybrid approach of constructing credit scoring model with better classification accuracy and interpretable capability than all the existing credit scoring models.
2. CART can be treated as a simple feature selection method to extract influential variables, those chosen variables can be further explained by credit analysts.
3. This study also presents an extensive comparison among existing credit scoring models constructed by various data mining algorithms. Loan companies can employ the proposed hybrid model to establish their own credit scoring models.
4. From the practical point of view, the hybrid GMDH model can be easily applied in spread sheet files such as Excel. This merit can be very helpful to credit analysts of loan companies.



References

- [1] Alexy G. Ivakhnenko, and Hema R. Madala. (1994). *Inductive learning algorithms for complex systems modeling*, CRC press, Inc.
- [2] Altman, Edward, “Discriminant analysis and the prediction of corporate bankruptcy”, *The Journal of Finance*, Vol. 23, No. 4, pp 589-609, 1968
- [3] Armingier, G., Enache, D., and Bonne, T. (1997). Analyzing Credit Risk Data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks, *Computational Statistics*, 12, 293-310.
- [4] Breiman, L., Friedman, J. R. Olshen, and C. Stone. (1984). *Classification and Regression Trees*, Wadsworth, Pacific Grove, California, USA.
- [5] Brill, J. (1998). The importance of credit scoring models in improving cash flow and collections, *Business Credit*, 1, 16-17.
- [6] Desay, V. S., Crook, J. N., and Overstreet Jr., G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment, *European Journal of Operational Research*, 95, 24-37.
- [7] Donald F, Specht. (1991). A General Regression Neural Network, *IEEE Trans. Neural Networks*, vol. 2, no. 6, November, 568-576.
- [8] Fausett, L. (1994). *Fundamentals of neural networks, architectures, algorithms, and applications*, USA: Prentice Hall International, Inc.
- [9] Glorfeld, L. W., and Hardgrave, B. C. (1996). An improved method for developing neural networks: the case of evaluating commercial loan creditworthiness, *Computers Ops Res*, Vol. 23, No. 10, pp.933-944.

- [10] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. New York: Springer-Verlag New York. Inc.
- [11] Huang, Z., Chen, H. C., Hsu, C. J., Chen, W. H., and Wu, S. S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study, *Decision Support Systems*, 37, 543-558.
- [12] Lee, K. C., Han, I., and Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions, *Decision Support Systems*, 18, 63-72.
- [13] Lee, T. S., Chiu, C. C., Lu, C. J., and Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique, *Expert Systems With Applications*, 23, 245-254.
- [14] Markham, I. S., and Ragsdale, C. T. (1995). Combining neural networks and statistical predictions to solve the classification problem in discriminant analysis, *Decision Sciences*, 26, 2, 229-242.
- [15] Olmeda, I., and Fernandez, E. (1997). Hybrid classifiers for financial multicriteria decision making: the case of bankruptcy prediction, *Computational Economics*, 10, 317-335.
- [16] Piramuthu, S., Ragavan, H., and Shaw, M. J. (1998). Using feature construction to improve the performance of neural networks, *Management Science*, Vol. 44, No. 3, 416-430.
- [17] Sharma, S. (1996). *Applied multivariate techniques*, New York: John Wiley & Sons, Inc.
- [18] Tam, K. Y. and Kiang, M. Y. (1992). Managerial applications of neural networks:

the case of bank failure predictions, *Management Science*, Vol. **38**, No. 7, 926-947.

[19] Vellido, A., Lisboa, P. J. G., and Vaughan, J. (1999). Neural networks in business: a survey of applications (1992-1998), *Expert Systems With Applications*, 17, 51-70.

[20] West, D. (2000). Neural network credit scoring models, *Computers & Operations search*, 27, 1131-1152.

[21] Zhang, G. P. (2000). Neural networks for classification: a survey, *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, Vol. **30**, No. 4, 51-462.



Appendix

