A Hybrid Approach to Classification

Venugopala Rao Manneni[1], Naveen Kumar Boiroju[2], M. Krishna Reddy[3]

[1]Department of Statistics, Osmania University Campus,
Hyderabad 500 007, email: venugopal.manneni@gmail.com.

[2]Department of Statistics, Osmania University Campus,
Hyderabad 500 007, email: nanibyrozu@gmail.com.

[3]Department of Statistics, Osmania University Campus,
Hyderabad 500 007, email: reddymk54@gmail.com.

Abstract

Classification of credit risk is an important task and widely studied topic since it can have significant impact on bank lending decisions and profitability. In this paper, classification of credit risk using decision trees, feedforward neural networks and a hybrid approach of combining decision tree and neural networks is discussed. We identify the key factors for the classification of credit risk using decision tree method and use these identified factors as inputs for the feed forward neural network for predicting the good or bad credit risk under the proposed approach. The proposed hybrid method compared with the decision trees and feedforward neural networks using the classification accuracy and misclassification matrix.

Keywords- *Classification, Decision Tree, Feedforward Neural Networks, Credit risk.*

---

1. Introduction

A loan officer at a bank wants to be able to identify characteristics that are indicative of people who are likely to default on loans, and then use those characteristics to discriminate between good and bad credit risks. Credit risk is the primary financial risk in the banking system and exists in virtually all income-producing activities. How a bank selects and manages its credit risk is critically important to its performance over time; indeed, capital depletion through loan losses has been the proximate cause of most institution failures. Identifying and rating credit risk is the essential first step in managing it effectively. Well-managed credit risk rating systems promote bank safety and soundness by facilitating informed decision making. Decision-making problems in credit evaluation and its risk measurement are very important and difficult tasks for commercial banks and financial institutions due to the high level of risk associated with wrong decisions [1].

Classification is one of the most frequently encountered decision making tasks of human activity. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes related to that object. Many problems in business, science, industry, and medicine can be treated as classification problems. Examples include bankruptcy prediction, credit scoring, medical diagnosis, quality control, handwritten character recognition, and speech recognition [2].

One of the most important and widely used classification models is decision tree. A brief introduction and its applicability discussed in Section 2. Recently, artificial neural networks have been extensively studied and used in classification problems. With artificial neural networks, there is no need to specify a particular model form; rather, the model is adaptively formed based on the features presented from the data. This data-driven approach is suitable for many empirical data sets where no theoretical guidance is available to suggest an appropriate data generating process. A brief introduction of classification using neural networks is presented in Section 3. In Section 4, we propose a hybrid approach to classification using both decision tree and neural networks models. Final conclusions presented in the Section 5.

2. Classification Techniques

The theory of statistical classification deals with the problem of assigning one or more individuals to one of several possible groups or populations on the basis of a set of characteristics observed on them. Generally speaking, classification is the action of assigning an object to a category according to the characteristics of the object. In data mining, classification refers to the task of analysing a set of pre-classified data objects to learn a model (or a function) that can be used to classify an unseen data object into one of several predefined classes. A data object, referred to as an example, is described by a set of attributes or variables. One of the attributes describes the class that an example belongs to and is thus called the class attribute or class variable. Other attributes are often called independent or predictor attributes (or variables). The set of examples used to learn the classification model is called the training data set. Classification belongs to the category of supervised learning, distinguished from unsupervised learning. In supervised learning, the training data consists of

pairs of input data (typically vectors), and desired outputs, while in unsupervised learning there is no a priori output [2].

Classification has been studied in statistics and machine learning. In statistics, classification is also referred to as discrimination. Early work on classification focused on discriminant analysis, which constructs a set of discriminant functions, such as linear functions of the predictor variables, based on a set of training examples to discriminate among the groups defined by the class variable. Modern studies explore more flexible classes of models, such as providing an estimate of the joint distribution of the features within each class (e.g. Bayesian classification), classifying an example based on distances in the feature space (e.g. the k-nearest neighbour method), and constructing a classification tree that classifies examples based on tests on one or more predictor variables (i.e., classification tree analysis) [2] [3].

2.1 Decision Trees

Traditional statistical prediction methods (for example, regression, logistic regression or discriminant analysis) involve fitting a model to data, evaluating fit and estimating parameters that are later used in a prediction equation. In the field of machine learning, attention has more focused on generating classification expressions that are easily understood by humans. The most popular machine learning technique is decision tree learning, which learns the same tree structure as classification trees but uses different criteria during the learning process. The technique was developed in parallel with the classification tree analysis in statistics [4] [5].

Decision tree or rule induction models take a different approach. They successively partition a data set based on the relationships between predictor variables and a target (outcome) variable. When successful, the resulting tree or rules indicate which predictor variables are most strongly related to the target variable. They also find subgroups that have concentrations of cases with desired characteristics. Decision trees represent a set of decisions. These decisions generate rules for classification of a dataset using the statistical criterion: entropy, information gain, Gini index, chi-square test, measurement error, classification rate, etc.

Chi-square automatic interaction detection (CHAID) is a heuristic decision tree method, which examines the relationship between many categorical or discrete predictor variables and a categorical target or outcome measure. It provides a summary diagram depicting the predictor categories that make the greatest difference in the desired outcome. This summary diagram can be searched to locate the sub-groups with the highest percentages on the target outcome category, or a gains table can be examined. The greatest strength of classification tree approach is its transparency and ease of deployment and easy to understand. It is designed to identify synergetic interactions when it compared to the discriminant and logistic regression. The weaknesses of classification trees are they are data hungry and can take large amounts of time for model building.

2.2 Neural Networks

Neural networks, also referred to as artificial neural networks, are studied to simulate the human brain although brains are much more complex than any artificial neural network developed so far. A neural network is composed of a few layers of interconnected computing units (neurons or nodes). Each unit computes a simple function. The inputs of the units in one layer are the outputs of the units in the previous layer. Each connection between units is associated with a weight. Parallel computing can be performed among the units in each layer. The units in the first layer take input and are called the input units. The units in the last layer produce the output of the networks and are called the output units. When the network is in operation, a value is applied to each input unit, which then passes it's given value to the connections leading out from it, and on each connection the value is multiplied by the weight associated with that connection. Each unit in the next layer then receives a value which is the sum of the values produced by the connections leading into it, and in each unit a simple computation is performed on the value - a sigmoid function is typical. This process is then repeated, with the results being passed through subsequent layers of nodes until the output nodes are reached. Neural networks can be used for both regression and classification. To model a classification function, we can use one output unit per class. An example can be classified into the class corresponding to the output unit with the largest output value. Neural networks differ in the way in which the neurons are connected, in the way the neurons process their input, and in the propagation and learning methods used. Learning a neural network is usually restricted to modifying the weights based on the training data; the structure

of the initial network is usually left unchanged during the learning process. A typical network structure is the multilayer feed-forward neural network, in which none of the connections cycles back to a unit of a previous layer. The most widely used method for training a Feedforward neural network is backpropagation.

Neural networks have emerged as an important tool for classification. The recent vast research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods. The advantage of neural networks lies in the following theoretical aspects. First, neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. Second, they are universal functional approximators in that neural networks can approximate any function with arbitrary accuracy. Since any classification procedure seeks a functional relationship between the group membership and the attributes of the object, accurate identification of this underlying function is doubtlessly important. Third, neural networks are nonlinear models, which makes them flexible in modelling real world complex relationships. Finally, neural networks are able to estimate the posterior probability, which provides the basis for establishing classification rule and performing statistical analysis [6] [7].

3. Proposed Method

The decision tree is one of the most popular classification algorithms in current use in data mining and machine learning. A decision tree is a tree structure, which classifies instances by sorting them down the tree from the root node to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. Depending on whether the test is validated or not, we will descend on one branch or the other of the tree. Each leaf of the tree indicates a solution, i.e. a label, or class (several leaves can correspond to the same class). In order to train such a tree, at every step, we try to separate (segment) the training data, as neatly as possible, into two groups. This requires on the one hand selecting an attribute and then in the other hand choosing a criterion on this attribute. A decision tree does not provide a mathematical model for assigning a new case into existing groups, where as it does the same thing with the nodes.

If we are only interested in the best possible classification accuracy, it might be difficult or impossible to find a single classifier that performs well as a good ensemble of classifiers. We will integrate two or more methods based on merits of the methods to improve the classification accuracy as well as better interpretability and prediction. As we know that the decision tree does not provide the mathematical model for classification of new object into the defined groups, but it gives the suitable predictors for the classification. Whereas neural networks provides the mathematical model in the form of weight matrices, but it has the problem of selection of necessary input variables for the classification.

We will use decision tree method to reduce the dimensionality and we can find only significant explanatory variables as well as interactions for a classification problem. And reduced dimensionality used for classification using neural networks in the proposed method. An attempt is made to combine both decision trees and neural networks to get a mathematical model for classification with necessary variables that are identified using decision tree. This procedure explained in detail with an illustration in the next section.

4. Empirical Study

A publicly available data set known as the German credit data [3] contains observations on 20 variables for 1000 past applicants for credit. In addition, the resulting credit rating (good or bad) for each applicant was recorded. The objective is to develop a credit classification rule that can be used to determine if a new applicant is a good credit risk or a bad credit risk based on values for one or more of the 20 explanatory variables namely, 1)Status of existing checking account, 2) Duration in month, 3) Credit history, 4) Purpose, 5) Credit amount, 6) Savings account/bonds, 7) Present employment since, 8) Installment rate in percentage of disposable income, 9) Personal status and sex, 10) Other debtors / guarantors, 11) Present residence since, 12) Property, 13) Age in years, 14) Other installment plans , 15) Housing, 16) Number of existing credits at this bank, 17) Job, 18) Number of people being liable to provide maintenance for, 19) Telephone and 20) Foreign worker. Essentially, then we must develop a function of several variables that allows us to classify a new applicant into one of two categories Good or Bad credit risk [3].

We will develop a classification procedure using three approaches: decision tree, Feedforward neural networks and a hybrid approach of combining the decision tree and neural networks approaches with the help of SPSS 17 software. We will compare the performance of the three approaches on a validation (Holdout) data set. We use approximately 90% of data for modelling and remaining 10% of data (Holdout data) for validation of the model. The design and training procedures for each classification method employed are the following.

4.1 Classification using Decision Trees

Classification of good or bad credit risk using decision tree technique is discussed in this section. CHAID method is used to construct a decision tree for the classification of good or bad credit risk based on all the 20 explanatory variables. The decision tree contains the total 15 nodes and resulting terminal nodes are 9. CHAID analysis includes the following independent variables with minimum classification risk are: 1) Status of existing checking account, 2) Credit history, 3) Property, 4) Personal status and sex, and 5) Other installment plans. The resultant classification matrix is presented in Table 1.

Table 1: Classification accuracy using CHAID analysis

Sample	Observed	Predicted		
		Good	Bad	Percent Correct
Training	Good	522	107	83.00%
	Bad	124	146	54.10%
	Overall Percentage	71.90%	28.10%	74.30%
Holdout	Good	55	16	77.50%
	Bad	12	18	60.00%
	Overall Percentage	66.30%	33.70%	72.30%

From the above results it can be observed that the present CHAID algorithm classifies 74% of the cases correctly in training sample and it classify 72% of the cases correctly in holdout set with five independent variables only.

4.2 Feed Forward Neural Networks

A feed forward neural network is used to classify the good or bad credit risk based on all the 20 explanatory variables. We divide the total sample into training, testing and holdout samples. Partition of the data sets is presented in the following table.

Table 2. Case Processing Summary

		N	Percent
Sample	Training	688	68.80%
	Testing	210	21.00%
	Holdout	102	10.20%
Total		1000	100.00%

The feed forward neural network model is a three layer feed forward neural network and it consists of an input layer, one hidden layer and one output layer. Total number of input neurons needed in this model is 62 (for 20 input variables and a bias unit), number of hidden neurons is 2 (including bias) and two output neurons represents the good or bad credit risk. A backpropagation learning method is used to train the network and the following results are obtained.

Table 3. Classification matrix using Neural Networks

Sample	Observed	Predicted		
		Good	Bad	Percent Correct
Training	Good	429	51	89.40%
	Bad	86	122	58.70%
	Overall Percent	74.90%	25.10%	80.10%
Testing	Good	130	26	83.30%
	Bad	26	28	51.90%
	Overall Percent	74.30%	25.70%	75.20%
Holdout	Good	59	5	92.20%
	Bad	21	17	44.70%
	Overall Percent	78.40%	21.60%	74.50%

From the above results it can be observed that the present feed forward neural networks classifies 80% of the cases correctly in training sample, 75% of the cases correctly classified in testing set and it classify 74.5% of the cases correctly in holdout set.

4.3 Hybrid Approach

In this method, the resulting variables under decision tree method are used to classify the good or bad credit risk using Feedforward neural networks. The key variables, X_1= Status of existing checking account, X_2= Credit history, X3 = Property, X4 = Personal status and sex, and X5= Other installment plans are taken as inputs for the construction of neural networks model to classify the credit risk. We divide the total sample into training, testing and holdout samples. Partition of the data sets is presented in Table 4. The feed forward neural network model is a three layer feed forward neural network and it consists of an input layer, one hidden layer and one output layer. Input layer consists of the 21 units (including bias unit) for each level of the 5 independent variables identified in decision tree method. Number of hidden neurons is 2 (including bias) and two output neurons represents the good or bad credit risk. A backpropagation learning method is used to train the network.

Table 4. Case Processing Summary

		N	Percent
Sample	Training	711	71.10%
	Testing	191	19.10%
	Holdout	98	9.80%
Total		1000	100.00%

The below Table 5 displays the coefficient estimates that show the relationship between the units in a given layer to the units in the following layer. The synaptic weights are based on the training sample even if the active data set is portioned into training, testing and holdout data.

Table 5. Synaptic weights of the Feedforward neural network

Predictor		Predicted		
		Hidden Layer 1	Output Layer	
		H(1:1)	[Good]	[Bad]
Input Layer	(Bias)	-0.381		
	[X1=1]	-1.155		
	[X1=2]	-0.573		
	[X1=3]	0.184		
	[X1=4]	0.81		
	[X2=1]	-0.144		
	[X2=2]	-0.468		
	[X2=3]	0.41		
	[X2=4]	0.134		
	[X2=5]	0.734		
	[X3=1]	0.036		
	[X3=2]	-0.082		
	[X3=3]	0.449		
	[X3=4]	0.22		
	[X4=1]	0.57		
	[X4=2]	-0.003		
	[X4=3]	-0.37		
	[X4=4]	-0.473		
	[X5=1]	-0.17		
	[X5=2]	-0.234		
	[X5=3]	0.283		
Hidden Layer 1	(Bias)		0.66	0.34
	H(1:1)		0.283	-0.283

Hidden unit: Sum of the information under hidden neuron is

$$s = -0.381 - 1.155 * I(X_1 = 1) - 0.573 * I(X_1 = 2) + 0.184 * I(X_1 = 3) + 0.81 * I(X_1 = 4)$$
$$- 0.144 * I(X_2 = 1) - 0.468 * I(X_2 = 2) + 0.41 * I(X_2 = 3) + 0.134 * I(X_2 = 4) + 0.734 * I(X_2 = 5)$$
$$+ 0.036 * I(X_3 = 1) - 0.082 * I(X_3 = 2) + 0.449 * I(X_3 = 3) + 0.22 * I(X_3 = 4)$$
$$+ 0.57 * I(X_4 = 1) - 0.003 * I(X_4 = 2) - 0.37 * I(X_4 = 3) - 0.473 * I(X_4 = 4)$$
$$- 0.17 * I(X_5 = 1) - 0.234 * I(X_5 = 2) + 0.283 * I(X_5 = 3).$$

where $I(.)$ *is an indicator function*

And hidden output is $H_{(1:1)} = Tanh(s)$.

Output neuron presents the probability of good or bad credit risk and is given by

$$P(Good\,Credit\,Risk) = 0.66 + 0.283 * H_{(1:1)}$$

$$P(Bad\,Credit\,Risk) = 0.34 - 0.283 * H_{(1:1)}$$

.

We classify the new object into any one of the good or bad categories based highest predicted probability using the hybrid approach. Classification of credit risk using this hybrid technique yields the following results.

Table 6. Classification matrix using hybrid approach

Sample	Observed	Predicted		
		Good	Bad	Percent Correct
Training	Good	424	71	85.70%
	Bad	113	103	47.70%
	Overall Percent	75.50%	24.50%	74.10%
Testing	Good	125	15	89.30%
	Bad	31	20	39.20%
	Overall Percent	81.70%	18.30%	75.90%
Holdout	Good	62	3	95.40%
	Bad	15	18	54.50%
	Overall Percent	78.60%	21.40%	81.60%

From the above results it can be observed that the present hybrid approach classifies 74% of the cases correctly in training sample, 76% of the cases correctly classified in testing set and it classify 82% of the cases correctly in holdout set.

5. Conclusion

From the above study, it is clear that the proposed hybrid approach performing better than the decision trees and feedforward neural networks. And it is also observed that the hybrid approach can perform well with the limited number of variables and which makes the decision maker to concentrate on these key variables that are identified under decision tree. This hybrid approach gives a mathematical function for the classification of good or bad credit risk whereas the decision trees fails.

References

[1] Altman, Edward I. and Anthony Saunders, Credit Risk Management: Developments over the Last 20 Years, *Journal of Banking and Finance*, pp. 1721-1742, 1998.

[2] Hair, Jr. J.F., Anderson, R.E., Tatham, R.L. and Black, W.C., *Multivariate Data Analysis*, 5th Edition, Prentice-Hall International, USA, 1998 .

[3] Johnson, R.A and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th edition, Pearson education, 2002.

[4] Kartik, V., Venugopala Rao, M. and Sandhya, T., CHAID - A Classification Algorithm to Assess Child Development, *The SPSS Analyst*, Oct-Dec, 2008.

[5] Ravi Kumar, P. and Ravi, V., Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review, *European Journal of Operational Research*, 180, pp. 1–28, 2007.

[6] Wong, B.K. and Selvi, Y., Neural network applications in finance: A review and analysis of literature (1990-1996), *Information & Management*, Vol. 34, pp. 129-139, 1998.

[7] Zhang, G.P., Neural Networks for Classification: A Survey, *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, Vol. 30, pp. 451-461, 2000.