ISSN: 2229-6959 (ONLINE)
DOI: 10.21917/ijsc.2017.0204

FINANCIAL FORECASTING USING DECISION TREE (REPTree & C4.5) AND NEURAL NETWORKS (K*) FOR HANDLING THE MISSING VALUES

J. Jayanthi, Gurpreet Kaur and K. Suresh Joseph
[1,2]*School of Computer Science and Engineering, Lovely Professional University, India*
[3]*Department of Computer Science, Pondicherry University, India*

Abstract

Missing values are a widespread problem in data analysis. The purpose of this paper is to design a model to handle the missing values in predicting financial health of companies. Forecasting business failure is an important and challenge task for both academic researchers and business practitioners. In this study, we compare the classification of accuracy in decision tree methods (REP tree, C4.5) and with ANN method (K*) to handle the missing values.

Keywords:

Bankruptcy prediction, Missing values, Decision Tree (REPTree, C4.5), ANN (K*)

1. INTRODUCTION

There are some instances which may have missing attribute values hence making the data incomplete [3, 4]. There can be different reasons for which attribute values can be missing. In order to handle missing data, different methods are available like imputation method and case deletion methods. Imputation method involves the replacement of the missing values with estimates derived from applying statistical methods to the available data. Whereas deletion method involves the deletion of all the instances with missing values leading to the loss of useful information, thus introducing some bias in the data. Statistical analysis is highly obstructed with missing information. In addition to its representation in the loss of key data, it also introduces biased results in the analysis. In order to correct the problem of missing data, we need to employ a sound method of imputation which replaces missing [10] values with reasonable estimates.

Among different tools, bankruptcy prediction is a very good technique which helps different financial organizations and people in making different types of decisions based on different parameters including company's performance. Since 1960s, an extensive research has been done in the field of prediction of bankruptcy for financial firms. The tool of Bankruptcy aids the creditors, auditors, stakeholders and senior managers in identifying the problems at an early stage.

The appropriate people interfere in various business matters and issues at an early stage so that the cost of business failure is reduced to a greater extent. Based on these things, we can state the problem as: A given set of parameters will be there describing the present scenario of the company over a given period of time and based on this data, the prediction problem will be to find out the probability of the company to become bankrupt in the coming year.

This paper presents a performance comparison among different techniques used to handle the problem of missing values in predicting bankruptcy including various methods of decision trees and neural networks.

Different sections listed in this paper are: Section 2 is highlighting the appropriate background information in predicting bankruptcy. In section 3, related work is highlighted so that the analysis done in our results and experiments can be easily understood. Different methodologies used in this work are being explored in Section 4 including scope for further discussions.

2. BACKGROUND ON BANKRUPTCY PREDICTION

The problem is stated as: A given set of parameters will be there describing the present scenario of the company over a given period of time and based on this data, the prediction problem will be to find out the probability of the company to become bankrupt in the coming year. A huge number of algorithms exist which are used to construct classification model for bankruptcy prediction [1, 11] like statistical techniques , Case Based Reasoning, Neural Networks, Operational Research, Rough set, Evolutionary technique, Fuzzy Logic, Isotonic separation, Wavelet, Decision Tree and Hybridization techniques. In a variety of statistical techniques, different methods used are Linear Discriminant Analysis, Multivariate Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, Linear Probability Models (probit), Principal Component Analysis, Independent Component Analysis, Z score, Zeta model and Factor Analysis. Different techniques are employed in the field of artificial intelligence and included in the study belonging to different architectures of neural networks including Multi-Layer Perceptron (MLP), Radial Basis Function network (RBFN), and Cascade Correlation Neural network (CASCOR). Different techniques employed in the field of operational research include Linear Programming (LP), Data Envelopment Analysis (DEA) and Quadratic Programming (QP). Different techniques used in evolutionary algorithms include genetic algorithm, particle swarm optimization, ant colony optimization, artificial immune system etc.

3. LEARNING MODELS WITH DECISION TREE AND NEURAL NETWORK (REPTREE, C4.5 AND K*)

Data mining also known as Knowledge Discovery in Databases is the process in which meaningful patterns are discovered in huge databases. In addition to the extraction of meaningful patterns, it acts as an application providing significant and competitive advantages for making right decisions. Among different techniques of data mining which are most commonly used, decision trees provide the combined functions of

classification and prediction simultaneously. In a decision tree, nodes which are internal are considered as tests and the leaf nodes are treated as categories. For getting the correct output in context to the input pattern, filtering of the tests is done downwards through the tree. Decision trees have an advantage of not requiring any kind of statistical knowledge for prediction tasks. Input data sets are represented in the form of decision table and the missing values are represented in the decision table by "?". Decision trees are most commonly used techniques used for the process of decision making [6]. This process involves the construction of a decision tree having different branches and leaves which represents various factors with respect to a given situation. A decision tree in simple terms just behaves like a decision support tool. A graph of decisions similar to a tree having the possible outcomes including different factors like resource costs, results of events and utility. It can be considered as a method for displaying an algorithm. In an algorithm, following steps are followed basically:

- Base cases are evaluated.
- For each node representing an attribute, say, *x*, normalized information gain is calculated which is normalized and done by doing a split on *x*.
- After calculating gain for all attributes, assume that *x_best* is the best attribute having highest normalized information gain.
- Then, a decision node is created that will split on *x_best*.
- These steps are performed again in the form of recursion on the sub trees which are attained by performing a split on *x_best* and these nodes are added as the child nodes.

3.1 C4.5

C4.5 [2] algorithm is basically a top to down decision tree. During every stage, the most relevant and predictive attribute which is having highest information gain is found and based on this node, further nodes are split. Every node in the tree depicts a point of decision around some value of the attribute. In C4.5 algorithm, following considerations are made:

i. A set of pre-classified samples denoted by $PS = ps_1, ps_2,...$is taken as training data whereas the sample is denoted as a vector of features represented by $ps_i = x_1.....x_n$.

ii. The set of training data is then added with a class represented by Ci to which each sample belongs to where i can be 1 to *n*.

In this algorithm, the problem of missing values is also handled. For this, the type of the attribute value is checked first, if it is new, it is substituted with a value which is most commonly occurring known as mode. If the type is numerical, it is replaced with most common occurring value mode. Missing nominal attributes are assigned as label M and M is treated to be just like other attribute value. This method of handling missing values performs poorly in comparison with filling up with mean or mode.

3.1.1 Mathematical Model:

$$E(S) = \sum_{i=1}^{m} -P(f_i)\log_2 P(f_i)$$

$$G(S,A) = E(S) - \sum_{i=1}^{m} P(A_i)E(S_{Ai})$$

3.2 REPTree

This learning technique is based on C 4.5 algorithm and is comparatively fast method of learning [4] and this technique is capable to provide regression trees classification. A tree is made known as decision or regression tree on the basis of information gain or variance. The extra information which is not required is removed using a pruning technique called as reduced-error pruning. The attributes are evaluated to construct a decision tree with respect to a goal represented by an attribute which is quantitative in nature. In this method of constructing decision tree is done using the method of variance reduction in order to get a split which is balanced and hence resulting in reducing corrections in errors. These steps are performed recursively to get a more simplified and less error free tree as compared to the ones generated in the earlier steps till the time a given stopping condition is met.

3.2.1 Mathematical Model:

$$(Y) = -\sum_{i=1}^{m} P(Y = y_i)\log P(Y = y_i)$$

$$H(Y \mid X) = -\sum_{i=1}^{m} P(X = x_i)H(Y \mid X = x_i)$$

$$IG(Y;X) = H(Y) - H(Y \mid X)$$

3.3 K*

K* [5, 9] is an instance based learner in which a measure and examination of the performance on a range of problems is used. In instance based learning mechanisms, a set of already classified examples are taken and used to classify a new instance based on these examples. In order to handle the problem of missing values, a measure of entropic distance is used by the instance based classifier. Missing valued attributes are treated as a separate value and these are replaced with the average value.

3.3.1 Mathematical Model:

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

$$\text{Manhattan} = \sum_{i=1}^{k}|x_i - y_i|$$

$$\text{Minkowski} = \left(\sum_{i=1}^{k}\left(|x_i - y_i|\right)^q\right)^{1/q}$$

4. EXPERIMENTAL SETUP AND DISCUSSION

4.1 DATASET

The dataset used in this research is the bankruptcy data set (http://www.pietruszkiewicz.com) in literature. It includes 120 companies from a period of two consecutive years. Among the companies, 56 companies were bankrupted 2-5 years later. The Table.1 illustrates each company is described by 30 attributes.

Table.1. Attributes

Bankruptcy Dataset	
X1	Cash/current liabilities
X2	Cash/ total assets
X3	Current assets/ current liabilities
X4	Current assets/ total assets
X5	Working capital/ total assets
X6	Working capital/ sales
X7	Sales/ inventory
X8	Sales/ receivables
X9	Net profit/ total assets
X10	Net profit/ current assets
X11	Net profit/ sales
X12	Cross profit/ sales
X13	Net profit/ liabilities
X14	Net profit/ equity
X15	Net profit/ (equity + long term liabilities)
X16	Sales/ receivables
X17	Sales/ total assets
X18	Sales/ current assets
X19	(365* receivables)/sales
X20	Sales/total assets
X21	Liabilities / total income
X22	Current liabilities/ total income
X23	Receivable/ liabilities
X24	Net profit/sales
X25	Liabilities/ total assets
X26	Liabilities/ equity
X27	Long term liabilities/ equity
X28	Current liabilities/ equity
X29	EBIT(earnings before interests and taxes)/total assets
X30	Current assets/ sales

4.2 EVALUATION METRICS

Performance metrics were evaluated based on the classification of confusion matrix. Here *TP*, *TN*, *FP*, *FN* represent the usual notation for the matrix in terms of true and positive results from the classifier. Recall and precision measures are good indicators of the classifier performance. Accuracy refers to the total correct classification for the set regardless of type. F1 score quantifies the tradeoff between recall and precision and indicative of the performance of the overall algorithm.

- Recall $R = (TP/(TP+FN))$
- Precision $P = (TP/(TP+FP))$
- Accuracy= $(TP+TN)/(TP+FN+FP+TN)$
- F-score $F1=2(R*P/R+P)$
- True positive (*TP*) = the number of predicted positive cases that are actually positive.
- True negative (*TN*) = the number of predicted negative cases that are actually negative
- False positive (*FP*)= the number of predicted positive cases that are actually negative
- False negative (*FN*) = the number of predicted negative cases that are actually positive.

The Table.2 illustrates the confusion matrix for positive and negative tuples.

Table.2. Confusion Matrix

Confusion matrix - Predicted class			
Actual class		C1	C2
	C1	TP	FN
	C2	FP	TN

4.3 EXPERIMENTAL DESIGN

The designs in this paper were performed using libraries from Weka 3.7.4 [12] machine learning environment. A lot of studies used in Weka in classification task, for examples. Fifteen selected decision tree classifiers are used to build the classification models; this classifier was briefly described above.

Each classification method was used as it is in Weka environment which means that no additional parameter tuning was performed before or during classification performance comparison. As well as we evaluate AUC of the classification methods using, each test we used 10 fold cross validation. We summarize our machine learning work for bankruptcy prediction in three main stages. First stage is attribute selection, second is choosing appropriate predictor, and finally evaluation of model as shown in Fig.1.

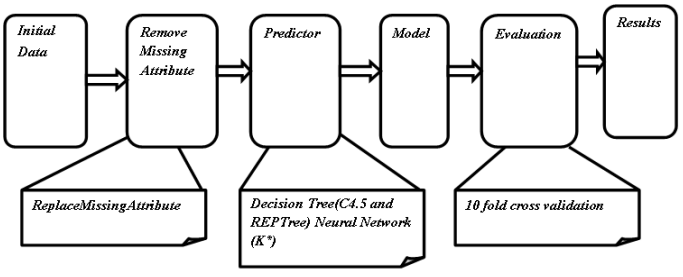

Fig.1. Bankruptcy Prediction Classification System

Table.3. Classification Result

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Class	Accuracy
C4.5 with missing values(original data set)	0.804	0.148	0.826	0.804	0.814	0.844	-1	82.5%
	0.852	0.196	0.832	0.852	0.842	0.844	1	
C4.5 handled missing values	0.795	0.148	0.824	0.795	0.809	0.831	-1	82.9%
	0.852	0.205	0.826	0.852	0.838	0.831	1	
REPTree with missing values (original data set)	0.804	0.172	0.804	0.804	0.804	0.845	-1	81.6%
	0.828	0.196	0.828	0.828	0.828	0.845	1	
REPTree handled missing values	0.821	0.172	0.807	0.821	0.814	0.847	-1	82.5%
	0.828	0.179	0.841	0.828	0.835	0.847	1	
K* with missing values (original data set)	0.786	0.164	0.807	0.786	0.796	0.88	-1	79.1%
	0.836	0.214	0.817	0.836	0.826	0.879	1	
K* handled missing values	0.768	0.188	0.782	0.768	0.775	0.877	-1	81.2%
	0.813	0.232	0.8	0.813	0.806	0.875	1	

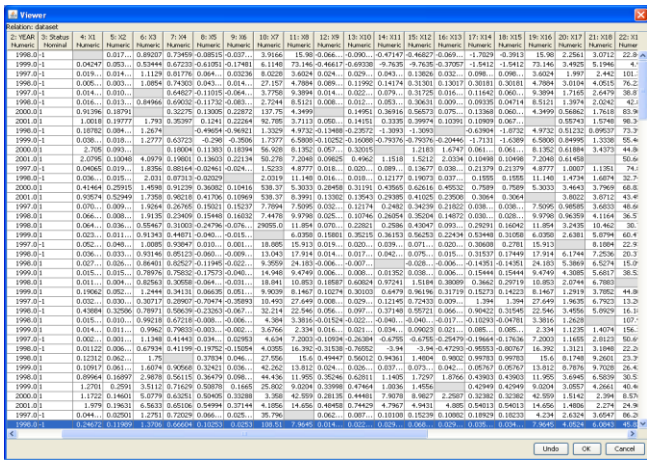

Fig.2. Bankruptcy Dataset (missing values)

4.4 RESULT AND ANALYSIS

We report the results from three methods that used to study the usefulness of the decision tree and neural network to predict bankruptcy. For each method we evaluated classification accuracy on the original dataset. We notice that C4.5 gave the higher accuracy on original data set (82.9%) with handled missing values. We observe that F1 score also higher in C4.5 (82.4 %) and Error Type I & II are better than other methods. The Table.3 summarized the accuracy of these methods. An overall view of the binary classifier performance is observed in Fig.2 which depicts the ROC curves demonstrating the performance

5. CONCLUSION

This experimental study compares classification performance of decision tree and ANN via using bankruptcy dataset for dealing with missing attributes. The algorithms taken into consideration are C4.5, REPTree and ANN (K*). From which we obtained experimental results conclude C4.5 is efficient rather than other two.

REFERENCES

[1] Wei-Yang Lin, Ya-Han Hu and Chih-Fong Tsai, "Machine Learning in Financial Crisis Prediction: A Survey", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 42, No. 4, pp. 421-436, 2011

[2] Damrongrit Setsirichok, Theera Piroonratana, Waranyu Wongsere, Touchpong Usavanarong, Nuttawut Paulkhaolarn, Chompunut Kanjanakorn, Monchan Sirikong, Chanin Limwongse and Nachol Chaiyaratana, "Classification of Complete Blood Count and Haemoglobin Typing Data by a C4.5 Decision Tree, A Naive Bayes Classifier and Multilayer Perception for Thalassaemica Screening", *Biomedical Signal Processing and Control*, Vol. 7, No. 2, pp. 202-212, 2012.

[3] Amita Karmaker and Stephen Kwek, "Incorporating An EM- Approach for Handling Missing Attribute Values in Decision Tree Induction", *Proceedings of IEEE 5th International Conference on Hybrid Intelligent Systems*, pp. 1-6, 2005

[4] Taghi M. Khoshgoftaar, Andres Follcco, Jason Van Hulse and Lofton Bullard, "Software Quality Imputation in the Presence of Noisy Data", *Proceedings of IEEE International Conference on Information Reuse and Integration*, pp. 484-489, 2006.

[5] John. G. Cleary and Leonard E. Trigg, "K* An Instance based Learner using Entropic Distance Measure", *Proceedings of International Conference on Machine Learning*, pp. 108-114, 1995.

[6] Elaze Zibanezhad, Daryush Foroghi and Amirthassan Monadjemi, "Applying Decision Tree to Predict Bankruptcy", *Proceedings of IEEE International*

Conference on Computer Science and Automation Engineering, Vol. 4, pp. 165-169, 2011.

[7] J. Jayanthi, K. Suresh Joseph and J. Vaishnavi, "Bankruptcy Prediction using SVM and Hybrid SVM Survey", *International Journal of Computer Applications*, Vol. 34, No. 7, pp. 39-45, 2011.

[8] Qin Zheng and Jiang Yanhui, "Financial Distress Prediction based on Decision Tree Models", *Proceedings of IEEE International Conference on Service Operations and Logistics, and Informatics*, pp. 1-6, 2007.

[9] Ming-Hua Chen, "Pattern Recognition of Business Failure by Auto Associative Artificial Neural Networks in Considering the Missing Values", *Proceedings of IEEE International Computer Symposium*, pp. 711-715, 2010.

[10] Maytal Saar-Tsechansky and Foster Provost, "Handling Missing Values when Applying Classification Models", *Journal of Machine Learning Research*, Vol. 8, pp. 1625-1657, 2007.

[11] P. Ravi kumar and V. Ravi, "Bankruptcy Prediction in Banks and Firms Via Statistical and Intelligent Technique-A Review", *European Journal of Operational Research*, Vol. 180, No. 1, pp. 1-28 ,2007.

[12] Nikolaos Mallios, Elpiniki Papageorgion and Michael Samarinas, "Comparison of Machine Learning Technique using the WEKA Environment for Prostate Cancer Therapy Plan", *Proceedings of IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pp. 151-155, 2011.

[13] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students Performance", *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 6, pp. 63-70, 2011.

[14] Abdullah AL-Malaise, Areej Malibari and Mona Alkhozae, "Students Performance Prediction System using Multi Agent Data Mining Technique", *International Journal of Data Mining and Knowledge Management Process*, Vol. 4, No. 5, pp. 1-20, 2014.

[15] Kamal Bunkar, Rajessh Kumar, Umesh Kumar and Singhand Bhupendra Pandya, "Data Mining: Prediction for Performance Improvement of Graduate Students using Classification", *Proceedings of 9th International Conference on Wireless and Optical Communications Networks*, pp. 1-5, 2012.

[16] S. Venkata Krishna Kumar and S. Padmapriya, "An Efficient Recommender System for Predicting Study Track to Students using Data Mining Techniques", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, No. 9, pp. 7996-7999, 2014.

[17] Dorina Kabakchieva, "Predicting Student Performance by using Data Mining Methods for Classification", *Cybernetics and Information Technologies*, Vol. 13, No. 1, pp. 61-72, 2013.

[18] Bashir Khan, Malik Sikandar Hayat Khiyal and Muhammad Daud Khattak, "Final Grade Prediction of Secondary School Student using Decision Tree", *International Journal of Computer Applications*, Vol. 115, No. 21, pp. 32-36, 2015.

[19] G. Naga Raja Prasad and A. Vinaya Babu, "Mining Previous Marks Data to Predict Students Performance in Their Final Year Examinations", *International Journal of Engineering Research and Technology*, Vol. 2, No. 2, pp. 1-4, 2013

[20] Jyoti Namdeo and Naveenkumar Jayakumar, "Predicting Students Performance using Data Mining", *International Journal of Advance Research in Computer Science and Management Studies*, Vol. 2, No. 2, pp. 367-373, 2014.

[21] Md. Hedayetul Islam Shovon and Mahfuza Haque, "Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, No. 7, pp. 353-355, 2012