The Comparative of Attribute Selection Techniques between CFS and Consistency by Using ANFIS for Thai Enterprises Bankruptcy Prediction

Kulthon Kasemsan* and Wonlop Buachoom**

Faculty of Information Technology, Rangsit University, Pathumtani, Thailand
kulthonkasemsan@yahoo.com*, bpolnow@gmail.com**

ABSTRACT – This paper presents the comparison of attribute selection techniques between CFS and Consistency for seeking better technique which could appropriately associate with ANFIS. Better model will be used for predicting business bankruptcy in Thai enterprises. According the objective of this study, there are two prediction models, CFS-ANFIS and Consistency-ANFIS. Type 1 error from estimation, which effect to stakeholders' decision making, is used for consider each model. The result indicates that estimation error rates obtained from CFS-ANFIS are lower than the error rates obtained from Consistency-ANFIS.

KEY WORDS – Attribute Selection, ANFIS, Bankruptcy Prediction

1. Introduction

Creditor and investor have an interest in assessing the financial position of a business and it potential for bankruptcy. They need to assess and predict the risk of non recovery of loan or investment before the extent of it. Then financial statement analysis is used for look at past and actual firm's financial ratios to predict its future situation [1].

According using of financial ratios for consider status of business, there were several prior researches which studied about the using of financial ratio for predicting bankruptcy or estimating the failure of business. Either statistical technique [2] or data mining technique was used to support the study's analysis. Furthermore, data mining technique for analytical was developed to the stage of Neural Network [1] and Neuro Fuzzy [3]. In case of Thailand, most papers, which studied about the business failure, employed the statistical technique to determine the result of the financial distress status of the business unit. Unfortunately, there were barely has any evidences which predicted the unstable status from the data mining forecasting ability.

Thus, this paper tried to create the bankruptcy forecasting model by using financial ratios as the input data in the data mining technique. In addition, the Adaptive Neuro Fuzzy Inference System: ANFIS was used to build up the failure estimating model. According to the result of previous studies pointed out that the integration between Fuzzy Logic and Neural Network had a good effect to construct the forecasting model, because the finalized hybrid model was combined with the advantages of both techniques together [4]. Furthermore, there was strongly supported by prior researches that found out the level of forecasting error from the using of Neuro Fuzzy was lower than the using of only Neural Network to create the estimating model [3].

With this paper, there are a lot of financial ratios which are used to simulate the forecasting model. So, this paper has to perform the attribute selection technique for this action. According to the previous papers, some attribute selection techniques could help reducing the error of business failure forecasting model [1, 3]. Simultaneously, many papers pointed out that the attribute selection technique would decrease size of input data [5]. In addition, this technique also increased the algorithm's performance and learning ability [5, 6]. Moreover, the accuracy rate of algorithm processing would be increased by using this technique [7]. However, there were benefits from attribute selection techniques for algorithm efficiency development, there was rarely technique which suitable for every situations [5]. To adapt model for any situations, the users have to join the advantages of several techniques together in the algorithm process.

Therefore, the objective of this paper is to compare the selection technique between the Correlation-based Feature Selection: CFS and the consistency-based Subset Evaluation: Consistency for seeking better technique which could appropriately associate with ANFIS.

2. Research Framework

2.1 Attribute Selection

The attribute selection is the selecting procedure, which selects M attributes as the subset from data set N attributes, to reduce a variety of relevant attribute and to ensure that the selected attributes have sufficient quality for processing [4]. Obviously, this selecting technique can help cutting off the irrelevant elements and the repeated data selecting of forecasting model [8]. The attribute selecting procedure has 2 core stages of selecting. The first stage of attribute selection is the attribute searching. This searching method is to seek out the relevant attributes. This paper uses the Forward Selection method as a selecting tool, which will set up the ordinary node path way to run the Information Gain continually. To illustrate, at the 1st round, it will select a single attribute as the representative of every attributes for calculating Information Gain of the whole attributes. According to this process, if one of the attribute has a highly gain, this attribute will be chosen to be a primary node for repeating the next loop [5]. The equation of Information Gain is as follow:

$$Gain(A) = I(S_1, S_2, ..., S_n) - E(A) \qquad (1)$$

From the (1) equation, A is the attribute which uses to calculate the Information Gain or Gain (A). This Information Gain points out that the selected attribute should be appropriately used to classify data among other attribute's gain.

Secondly, the next stage of attribute selection is the evaluating set of attribute from searching process. This stage is used to evaluate the suitability of the set of attribute. If the selected attribute presents the appropriated result, the system will definitely stop to select the attribute. With this stage, this paper focuses on the comparative of attribute selections between the CFS and Consistency. The process of CFS is to find out group of attribute. The set of attributes is evaluated by relationship between selected set of attributes, which uses for classifying type of data, and level of inter-related relationship of the set of attributes. In addition, the high relation with data classification and reversed internal relation the set of attributes is obtained the high point. The equation is used to score set of attributes, which is valued by Fayyad and Irani [5, 6], can be shown as follow:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (2)$$

According to the Fayyad and Irani's formula, $Merit_s$ is the set of attributes S which includes selected k attributes. $\overline{r_{cf}}$ is the average of the selected set of attributes which based on the type of data relation. $\overline{r_{ff}}$ is the average of the selected set of attributes which based on the internal relation.

The Consistency is used in choosing attributes, it focuses on the smallest number of attributes, which should reflect the most correctly result from forecasting. The Consistency equation was given by Lui and Setiono [5] is as follow:

$$Consistency_s = 1 - \frac{\sum_{i=0}^{j} |D_i| - |M_i|}{N} \qquad (3)$$

From the Consistency's calculation, $Consistency_s$ is group of attribute. J is the number of attributes in the group of attributes S. $|D_i|$ is the number of attributes that is grouped together. $|M_i|$ is the competitor of most of the class for the value of attribute that get together. And N is the total number of data.

2.2 ANFIS

ANFIS is the applied technique which uses for merging the Artificial Neuron Network and Fuzzy Logic together. With this technique, ANFIS can compensate the limitation of one procedure with the advantage of another one. In addition, the ability to learn and to adapt from the data characteristic is the advantage of Artificial Neuron Network, but this procedure has a critical constraint about the explanation of in-depth learning which every computer programs cannot understand the learning dimensions as good as human can. According to ANFIS technique, this limitation can be compensated by the fundamental stage of Fuzzy Logic which is

developed from the if-then rule in the Crisp Logic decision to be the Fuzzy decision [7]. The ANFIS has many layer components in it [7, 9, 10]. And the figure of layers components is as follow:

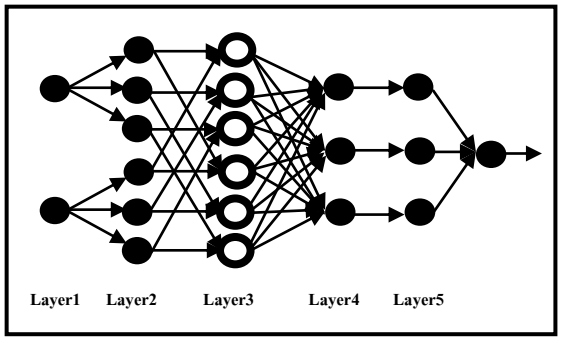

Figure 1: ANFIS Model

According to the ANFIS Model diagram, there are six layers of logics for this study. The first layer, input layer, replaces the data input with x which has n dimension. The first layer can be placed with the equation which $1 \leq i \leq n$ and θ is the ordinary number of testing data. The equation is as follow:

$$x_\theta = \left[x_{1\theta}, x_{2\theta}, ...x_{i\theta}\right] \tag{4}$$

Secondly, Fuzzification layer, the layer is used for Fuzzy valuation under Gaussian membership function. With this layer, μ is used for the second layer output. In addition, $1 \leq i \leq n$, $1 \leq j \leq R$, which n is the dimension of input components. R is amount of Fuzzy rule, and c and σ are the means and class interval of Gaussian membership function. Fuzzification equation is shown as:

$$y_{i,j}^{(\mu)} = e^{-\frac{\left(x_{i,\theta} - c_{i,j}\right)}{2\sigma_{i,j}^2}} \tag{5}$$

The third layer is the layer of Fuzzy rule. With this layer, the processing of this layer is to integrate the output of the second layer with Fuzzy rule from following equation, which Ru is the result of the third layer.

$$y_j^{(Ru)} = \prod_{i=1}^{n}\left(y_{i,j}^{\mu}\right) \tag{6}$$

The next layer is the Normalization layer. The output of this layer is generated from the division between the third output (numerator) and the summation of Fuzzy Rule (Denominator). The equation for this layer is as follow:

$$y_j^{(\bar{\mu})} = \frac{y_j^{(Ru)}}{\sum_{j=1}^{R} y_j^{(Ru)}} \tag{7}$$

After Normalization procedure, the pre-final layer is the Defuzzification layer. In this layer, the procedure brings the result of the last layer to multiply with the multiplier factor which is a parameter (k) from the Moore-Penrose Pseudo Inverse of a Matrix. In addition, the parameter (k) will have dimension $R \times (n+1)$. Thus, the formula of this layer is:

$$y_j^{(DF)} = y_j^{(\bar{\mu})} \bullet \left(\sum_{i=1}^{n}\left(k_{j,i} x_{i,\theta}\right) + k_{j,(n+1)}\right) \tag{8}$$

Finally, the layer is called the layer of Neuron Summarization or ANFIS output. This output is obviously appeared in the form of Fuzzy Sugeno function which can be calculated when the input (x_θ) and parameter $(\{k, c, \sigma\})$ are be set. The form of ANFIS output can be showed as follow:

$$TS\left(x_\theta, \{k, c, \sigma\}\right) = y_\theta^{(TS)} = \sum_{j=1}^{R} y_j^{(DF)} \tag{9}$$

3. Methodology

3.1 Data and Sample size

This paper uses the financial ratios for estimating the failure of business unit in case of Thailand totally amount 36 ratios [11, 12, 13]. The table 1 shows the lists of available ratios.

Sample size of this study is listed companies in the Stock Exchange of Thailand. These companies can be divided into two categories. One category consists of a set of fail companies, which has a Rehabilitation status: REHABCO under the Stock Exchange Commission's declaration. And other one consists of not fail companies. Furthermore, this study uses the data during the year 2005 to 2007 of 10 companies which have the REHABCO status and 99 companies which have normal status.

Table 1 List of available ratios

Ratio	Formula
R01	Cash/Current Liabilities
R02	Current Assets/Current Liabilities
R03	(Current Assets-Inventories)/Current Liabilities
R04	Sales/Average Accounts Receivable
R05	Cost of good sold/Average Inventories
R06	Sales/Working Capital
R07	Sales/Average Current Assets
R08	Sales/Average Fixed Assets
R09	Sales/Average Total Assets
R10	Sales/(Total Assets-Total Liabilities)
R11	Purchases/Average Accounts Payable
R12	Long-term Debt/(Long-term Debt+Stockholders' equity)
R13	Long-term Debt/Stockholders' equity
R14	Long-term Debt/Total Assets
R15	Total Liabilities/Total Assets
R16	Total Liabilities/Stockholders' equity
R17	Common Equity/Long-term Debt
R18	Stockholders' equity/Total Assets
R19	Gross Profit/Sales
R20	Operating Income/Sales
R21	Net Profit/Sales
R22	Operation Expenses/Sales
R23	Net Profit/Total Assets
R24	Net Profit/Stockholders' equity
R25	Net Profit/(Long-term Debt+Stockholders' equity)
R26	Net Profit/Dividend
R27	Operation Expenses/Interest
R28	Net Profit/Number of shares outstanding
R29	Dividend/Number of shares outstanding
R30	Stockholders' equity/Number of shares outstanding
R31	Dividend/Net Profit
R32	Cash from Operating/Current Liabilities
R33	Cash from Operating/Total Liabilities
R34	Cash from Operating/Number of shares outstanding
R35	Cash from Operating/Dividend
R36	Cash from Operating/Current Portion of Long-term Debt

3.2 Research methodology

Firstly, the collected data is divided into 2 sets which are the Training Set and the Test Set under the proportion 70 percent and 30 percent respectively. In the Training Set, there are 7 REHABCO companies and 70 normal companies. In contrast, the Testing Set has 3 REHABCO companies and 29 normal cases. The classified data is stated in the following table.

Table 2: Training set and test set details

Data Detail	Training Set	Test Set	Total
Separated Rate (%)	70	30	100
Fail Samples	7	3	10
Not Fail Samples	70	29	99
Total Samples	**77**	**32**	**109**
Fiancial Ratios	36	36	36
Collection Period (Year)	5	5	5
Total Data	**13,860**	**5,760**	**19,620**

After identified sets, this paper brings the Training Set to the attribute selection process by using CFS and Consistency methods. When the attribute selection process is complete, the forecasting failure model is setting up from ANFIS by using Matlab 2009. According to the result of the program calculation, there are 2 forecasting failure models, CFS-ANFIS and Consistency-ANFIS.

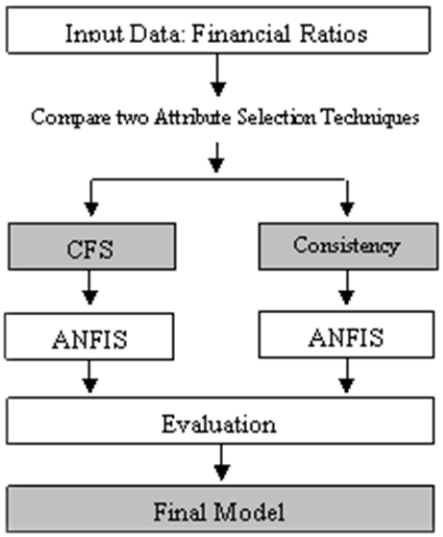

Figure 2: Framework of study

Both forecasting models are evaluated by another set. During the evaluating process, this paper toughly considers about type 1 error from the estimation of each forecasting model. This error type shows the negative result from the fact. To illustrate, the result points out that this company still on the normal status, in fact this firm encounters with the failure status [2]. With this error type, it enormously effects to the end-users' decision making.

4. Experimental Results

4.1 Actual Model

Generally, CFS-ANFIS and Consistency-ANFIS can exactly formulate the 2 different forecasting models for abnormal operating of business in Thailand. With the internal algorithm, ANFIS can learn and create the learning rules from input data and a target attribute is set up for each forecasting model. In addition, a target attribute is possibly occurred for 2 values which are 0 (encountering the failure situation) and 1 (normal status).

This study shows that CFS-ANFIS has 7 learning rules from input data, 25 selected attributes and 1 target attribute, but Consistency-ANFIS creates 4 learning rules form input data, 15 selected attribute and a additional target attribute.

4.2 Result Analysis

After generating forecasting model, the test set provides the error term of prediction for CFS-ANFIS for 6.25 percent which includes type 1 error equally to type 2 error at 3.13 percent. In contrast, the Consistency-ANFIS has the mistaken process for 9.38 percent. Surprisingly, all of the error which is found in the Consistency-ANFIS is type 1 error. On the other hand, it is shown that accuracy of failure estimation of CFS-ANFIS model is higher than Consistency-ANFIS model. In addition, type 1 error of CFS-ANFIS lower than Consistency-ANFIS. It clearly pointed that CFS-ANFIS better than Consistency-ANFIS for failure estimation in case of Thai enterprises.

Table 3: Error rate of models

Model	Overall Error		Error Type 1		Error Type 2	
	Samples	Rate	Samples	Rate	Samples	Rate
CFS-ANFIS	2	6.25%	1	3.13%	1	3.13%
Consistency-ANFIS	3	9.38%	3	9.38%	-	-

5. Conclusion

According to the objective of the study, this paper emphasizes on the comparison of the attribute selection techniques which are CFS and Consistency. Furthermore, these techniques can appropriately associate with ANFIS for estimating the failure situation in case of Thailand. Finally, CFS-ANFIS has lower error rate of prediction than Consistency-ANFIS approximately 3 percent. Therefore, in case of Thailand, this study recommends end-users to use the CFS-ANFIS for predicting bankruptcy or estimating the failure of business unit.

6. References

[1] Abdelwahed, T. and Amir E.M. " New Evolutionary Bankruptcy Forecasting Model Based on Genetic Algorithms and Neural Network". *Proceedings of the 17th IEEE International Conference on Tool with Artificial Intelligence*, 2005.

[2] Altman, Edward I. "Financial Ratios Discriminant Analysis and The Prediction of Corporate Bankruptcy". *The journal of Finance*. 4 (1968) : 589-609.

[3] Huang Fu-yuan. "A Genetic Fuzzy Neural Network for Bankruptcy Prediction in Chinese Corporations". *Proceeding of the 2008 International Conference on Risk Management & Engineering Management. IEEE Computer Society*, 2008.

[4] Tong Srikhacha. "Short-Term Prediction in Stock Price Using Hybrid Optimized Recursive Slope Filtering, Adaptive Moving Approach and Neurofuzzy Adaptive Learning". PhD thesis, Department of Information Technology, King Mongkut's Institute of Technology North Bangkok, Thailand, 2007.

[5] Borges, Helyane B. and Nievola, Julio C. "Attribute Selection Methods Comparison for Classification of Diffuse Large B-Cell Lymphoma". *Proceedings of the Fourth International Conference on Machine Learning and Applications. IEEE Computer Society*, 2005.

[6] Hall, Mark A. and Holmes, Geoffrey. "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining". *IEEE Trans on Knowledge and Data Engineering*. 15 (2003) : 1437-1447.

[7] Sukontip Wongpun and Anongnart Srivihok. "Comparison of Attribute Selection Techniques and Algorithms in Classifying Bad Behaviors of Vocational Education Students". *Proceedings of 2008 Second IEEE International Conference on Digital Ecosystems and Technologies. IEEE Computer Society*, 2008.

[8] Sukontip Wongpun. "Comparison of Attribute Selection Techniques and Algorithms in Classifying Mistaken Behaviors of Vocational Education Students". Master of Science thesis, Department of Computer Science, Kasetsart University, Thailand, 2008.

[9] Jang, J. S., Sun, C.T. and Mizutani, E. Neuro-Fuzzy and Soft Computing : A Computational Approach to Learning and Machine Intelligence. n.p. : Prentice-Hall, 1997.

[10] Jang, S. R. "ANFIS: Adaptive-Network-Based Fuzzy Inference System". *IEEE Trans. System, Man and Cybernetic*. 23 (1993) : 665-684.

[11] Gibson, Charles H. Financial Reporting and Analysis. Ohio : Thomson South-western, 2007.

[12] Stickney, Clyde P. Financial Reporting and Statement Analysis.Texas:The Dryden Press, 1996.

[13] Walsh Ciaran. Key Management Ratios The Clearest Guide to The Critical Numbers That Drive Your Business. Pearson Education, 2006.