Jurnal Teknologi

Full Paper

COMPARATIVE PERFORMANCE FOR PREDICTIVE MODELLING IN MOTOR INSURANCE CLAIMS

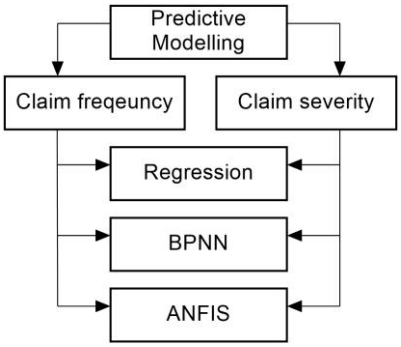
Zuriahati Mohd Yunos[a*], Siti Mariyam Shamsuddin[a], Razana Alwee[a], Noriszura Ismail[b], Roselina Salleh@Sallehuddin[a]

[a]Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[b]Pusat Pengajian Sains Matematik, Universiti Kebangsaan Malaysia, Selangor, Malaysia

Article history
Received
5 September 2016
Received in revised form
14 November 2016
Accepted
8 November 2016

*Corresponding author
zuriahati@utm.my

Graphical abstract



Abstract

The expected claim frequency and the expected claim severity are used in predictive modelling motor insurance claims. There are two categories of claims were considered, namely, third party property damage and own damage. Datasets from the year 2001 to 2003 are used to develop the predictive model. This paper proposes three different methods, namely, regression analysis, back propagation neural network and adaptive neuro fuzzy inference system to model claim frequency and claim severity as the two important elements in modelling the motor insurance claims. The experimental results showed that the back propagation neural network model produces more accurate as compared to regression analysis and adaptive neuro fuzzy inference system in predicting the claim frequency and claim severity. For both OD and TPPD claim, the results have shown the lowest MAPE with 0.2191 and 0.6515, and 0.2169 and 0.326, respectively.

Keywords: Predictive modelling, claim frequency, claim severity, regression, BPNN, ANFIS

1.0 INTRODUCTION

Predictive modelling in the insurance industry helps actuaries and other insurance analysis employing predictive models to enhance business operations that were previously using human expertise. Historically, actuaries performed their duties using pencil and paper before the advent of computers. Today, more advanced computing tools are available [19]. Predictive modelling has provided a set of instruments to the insurance companies for a variety of intentions from pricing to underwriting and claim handling [3]. Moreover, the influences of predictive modelling are also dependent on the quality of the data used to generate the model. Insurance is a unique type of agreement between the insurer or insurance company and the insured or client in which the insurers permit that upon the occurrence of specific events, whether to make a payment to clients or cover the specific costs. This research focuses to develop a predictive model for motor insurance claim by estimating the two important components, namely, claim frequency and claim severity [9, 2, 18, 25, 27, 30]. Claim frequency is defined as the number of claims per exposure unit, whereas claim severity is defined as the average claim cost per claim [18.]. The modelling of claim frequency and claim severity needed an information of exposure, number of claims and the amount of the claim (cost). The expected of claim frequency and claim severity can be calculated through a process of identifying grouping risk, which having the same characteristics is also known as risk classification.

Several studies have been carried out in modelling the motor insurance claims using statistical methods. For example, Ismail applied the normal, the exponential and the gamma regressions for fitting

claim severity data on Malaysian motor insurance claims, then Ismail used fitted negative binomial and generalized Poisson regressions for handling overdispersion in claim frequency data [18]; Morata suggested the bivariate Poisson regression to examine tariff ratemaking for two types of claims and rating factors [26]; and Freez proposed a hierarchical model for three components of claims, the negative binomial regression to estimate claim frequency, the multinomial logit model to predict type of claim, and the generalized beta distribution to estimate claim cost [12, 13] and others [7, 11, 20, 21]. Recently, several artificial intelligence (AI) approaches have been applied in modelling insurance claims. It is observed that the use of artificial neural network (ANN) and fuzzy logic (FL) especially adaptive neuro fuzzy inference system (ANFIS) has been increased during the last decade in the insurance field, for example the ANN model [2, 8, 11, 17, 22, 24, 33] and the FL model [5, 16, 29, 32]. ANFIS is one of the NeuroFuzzy techniques to solve a complex problem. ANFIS is the basis network architecture and its hybrid-learning rule is proposed by Roger Jang. Therefore, we present the ANFIS model as one of the benchmark model as this model is widely used in various fields such as bankruptcy prediction [34, 35], stock market [6, 15], financial [4] and others. Subsequently, a comparison between the forecast value and the actual value is executed by reducing the errors forecast of the predicted value and the actual value. The obtained results are then compared among the techniques. The remainder of this paper is organized as follows: Section 2 discusses the methodology used. Section 3 presents the datasets and model evaluation. Following are the results and discussion in Section 4 and the conclusion is provided in Section 5.

2.0 METHODOLGY

This section discusses the methodology used in this study, including the regression, backpopagation neural network (BPNN) and adaptive neuro fuzzy inference system (ANFIS). Figure 1 shows the flowchart in modelling the motor insurance claim for claim frequency and claim severity. In the flowchart the input factor has been chosen and then fed up into the methods. Then, each model process and produced the predicted value either the claim frequency or claim severity. The predicted output is compared with the actual output and if the predicted output doesn't meet the requirement (depends on the models requirement) the output is swapped back and rerun back by adjusting the parameters. Finally, if the predicted value meets the requirement, it will pass through the validation process.

2.1 Regression Model

A regression model is a linear model and it is a statistical model which describes the relationship between a dependent variable (*y*) and independent variables, $x_i, i = 1, 2..., n$.

The mathematical model is given in Equation (2.1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \quad (2.1)$$

where $\beta_0, \beta_1 ..... \beta_n$ are regression coefficients, and ε is the error due to variability in the observed responses.

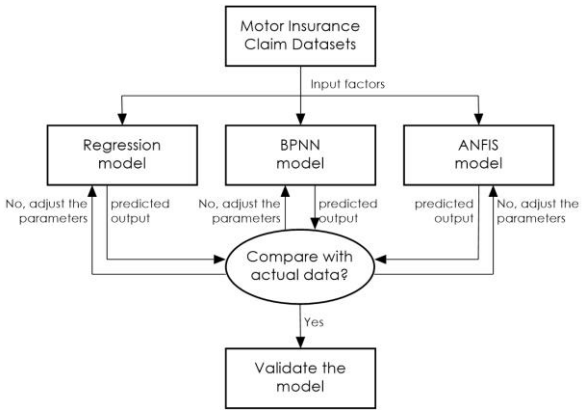


Figure 1 The predictive model flowchart for motor insurance claims

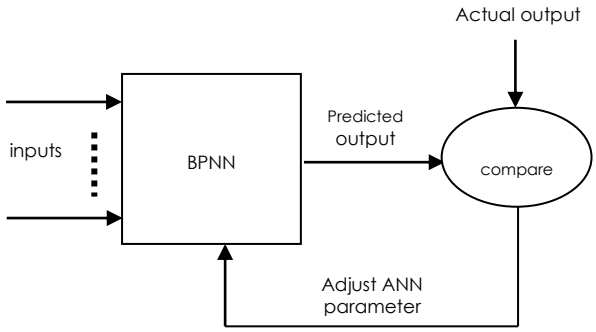


Figure 2 BPNN model diagram for motor insurance claims

2.2 BPNN Model

BPNN is one of the algorithms in ANN with a three-layer network structure and has been applied in many areas [1, 22, 36]. The related illustration of BPNN model diagram is given in Figure 2. The essential steps for designing BPNN model are summarized in Table 1. In particular, step 1 to step 4 are carried out in data pre-processing, where the raw data is scaled and normalized to an appropriate format to facilitate the predicting process. Step 5, which is the step that designs the ANN model, involves the determination of the following variables:

i. number of input nodes
ii. number of hidden layers and hidden nodes
iii. number of output nodes
iv. activation function

The normalization formula used is given in Equation (2.2) [23]:

$$X_{new} = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}}(D_{\max} - D_{\min}) + D_{\min} \quad (2.2)$$

where X_t is the value will be normalized, $X_{\min}$ is the minimum value of the statistic variable, and $X_{\max}$ is the maximum value of the statistic variable. $D_{\max}$ and $D_{\min}$ are the maximum and the minimum values needed for normalization. The values of $D_{\max} = 0.95$ and $D_{\min} = 0.0$ are set as it is the maximum value and minimum value for the normalized data. The normalization equation (2.2) is selected due to the range of the activation function utilized in the BPNN and also the real claim data value is a positive values and not negative values.

The BP algorithm involves two phases, the forward phase and the backward phase. In the forward phase, the activations are propagated from the input to the output layer, while in the backward phase, if the output pattern is different from the desired output, the error between actual and predicted values in the output layer is calculated and propagated backwards to modify the weights and bias values. The most popular error function used for the output layer is the mean sum squared error. The network is trained with a pre-defined stopping criterion; either the number of iterations has been reached or when the total sum of square errors is lowers than a pre-determined value. This is the core part of ANN. The BPNN architecture and parameters is shown in Table 2 and Table 3 describes the tested network structures.

Table 1 Steps in designing a BPNN model

Step 1	Variable selection
Step 2	Data collection
Step 3	Data normalization
Step 4	Data division: training and testing
Step 5	Determine the : - number of input nodes - number of hidden layers - number of hidden nodes - number of output nodes - activation function
Step 6	ANN training by applying BP algorithm : - set the learning rate and momentum - set the number of training iterations
Step 7	Model evaluation

Table 2 Summary of standard BPNN architecture and parameters chosen

Number of input nodes	4,5 and 6	
Number of hidden layer	1	
Number of hidden nodes	See Table 3 (tested network structures)	
Number of output nodes	1	
Learning rate	0.3	
Momentum	0.9	
Activation function	Input to hidden layer	sigmoid
	Hidden layer to output	sigmoid
Error performance	Mean square of error (MSE)	

Table 3 Tested network structures [a]

4-4-1	4-8-1	4-9-1
5-5-1	5-10-1	5-11-1
6-6-1	6-10-1	6-12-1

[a] Hidden nodes is based on *n*, *2n*, *2n +1* and *random*

2.3 ANFIS Model

ANFIS model is one of the neurofuzzy (NF) model. ANFIS model is the integration of fuzzy systems with neural networks. This model is based on Takagi and Sugeno model [31]. This learning method works similarly to that of neural networks. Figure 3 shows the basic diagram of ANFIS computation. Table 4 shows the results produced from fuzzy decision tree that consist number of input nodes, degree of membership, number of rules and different network structured. As a result, different network structures are constructed and a number of membership functions are tested to obtain the best network model with optimum features and suitable learning parameters. This information is crucial in developing the predictive model.

The training is based on 100 epochs and the error goal is set to 0. Hence, the stopping criterion is determined either when the epoch is complete or when the error goal is reached. For the learning algorithm, ANFIS has two learning cycles, namely; backpropagate and hybrid. The backpropagate learning is founded on the gradient descent method, while the hybrid learning is a combination of least-squares and backpropagation gradient descent method. The backpropagate learning is chosen for the ANFIS development.

3.0 DATA SET AND MODEL EVALUATION

This section describes the data set used and the model evaluation carried out in this study.

3.1 The Datasets

The datasets used is provided by Insurance Services Malaysia Berhad (ISM), which is based in the year 2001 until 2003, compiled from ten local insurance companies. The datasets used is provided by Insurance Services Malaysia Berhad (ISM), which is based for the year 2001 until 2003, compiled from ten local insurance companies. There are two types of claim data being considered; third party property damage (TPPD), and own damage (OD). In motor insurance, term of rating factors is also known as rating variables. It is also indicated as variables or features or inputs which are used to compute and predict the motor insurance claim such age, gender, vehicles age and others.

The data used in the experiments are first pre-processed by eliminating any missing values. After that, the datasets must be normalized to smooth out the data, resulting in better data generalization. Then the data is partitioned into two parts to obtain the training sets and testing sets. The percentage ratio between the training data and the testing data is 70%: 30% [36]. The inputs of training data and output for claim frequency and claim severity are shown in Table 1. For claim frequency, each data are used to determine the number of claims made by the insured (clients) and as claim severity the data are used to compute the amount claimed by the insured or amount paid by the insurer to insured.

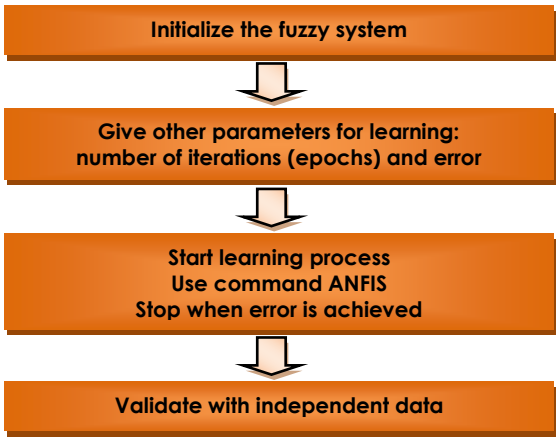


Figure 3 Basic diagram of ANFIS computation

4.0 RESULTS AND DISCUSSION

This study discusses the two models namely, claim frequency and claim severity and were tested on two category of claims which are TPPD and OD. The development of regression, BPNN and ANFIS model are done through trial-and-error with the aim to obtain the best predictive result for claim frequency and claim severity.

The criterion used to determine the best model is by looking at the lowest error value given by MSE, RMSE, MAE and MAPE [36]. The purpose is to measure the performance of the predictive models that has been developed. However, if the result produced by the four error measurement is inconsistent, then MAPE is chosen [14]. The best MAPE is chosen based on nearest value to zero. Furthermore, the relative performances of the predictive models are based on the MAPE for each model divided by the best model. The mathematical formulas for the error functions are:

$$MSE = \frac{1}{n}\sum^{n}(x-y)^{2} \quad (3.1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum^{n}(x-y)^{2}} \quad (3.2)$$

$$MAPE = \frac{(\sum^{n}|x-y|)}{\sum^{n}(x)} \quad (3.3)$$

$$MAE = \frac{\sum(|x-y|)}{n} \quad (3.4)$$

where n is the sample size of data, x is the actual data, y is the predicted data and | | = the absolute value. The relative performances of the predictive models are based on the MAPE for each model divided by the best model. For OD claims, the lowest MAPE were obtained by BPNN with 0.2191 and 0.6515 for claim frequency and claim severity. The same result happened when using TPPD claim, with 0.2169 and 0.3261. Tables 6 and 7 have clearly shown that BPNN model gives a promising result compared to regression and ANFIS models in claim frequency and claim severity for all claim types (**highlighted with bold italic**). The lowest result achieced by the BPNN based on the nearest value to zero. The results from this study also depict that the predicting performance produced by the BPNN model outperformed the regression and ANFIS models. The experimental result reveals that the number of input nodes and hidden nodes, as well as the parameters chosen influenced the predictive accuracy.

5.0 CONCLUSION

In this paper, we applied regression analysis, BPNN and ANFIS method as a learning tool for motor insurance claims in predictive modelling. It is concluded that the BPNN model is successful in modeling the Malaysian motor insurance claims by using several of network structures. Several factors that significantly influenced the performance of the BPNN have also been discussed, namely the network structure (number of input nodes and number of hidden nodes), data preprocessing, the parameters and the error measurement. The main advantage of using BPNN is that the model is capable of dealing with non-linear data. Outcomes of the evaluation, analysis show that regression model is unable to give feasible performances. This is due to the weaknesses in the

model itself resulted in by its incapableness in handling non-linear data. It is also demonstrated that the ANFIS model results are not encouraging compared to BPNN model. The predicted results can also be applied to classify risks and to compute insurance premium, and the proposed method can also be used for other claim types of insurance. For future working areas, it is suggested that the technique suggested in this study is hybridized with other techniques to improve the prediction performance.

Table 4 Input and output variables for TPPD and OD claim

Notation	Claim Frequency		Claim Severity	
	Input nodes	Output nodes	Input nodes	Output nodes
F1	Coverage		Coverage	
F2	Vehicle made		Vehicle made	
F3	Vehicle cc		Vehicle cc	
F4	Vehicle year	Claim frequency	Vehicle year	Claim severity
F5	Location		Location	
F6	Exposure		-	
F7	-		Number of claim	

Table 5 Results generate through fuzzy decision tree

Number of input nodes	Degree of membership	Number of rules (p^n)	Network structure
4	2	16	4-16-1
5	2	32	5-32-1
6	2	64	6-64-1

Table 6 Comparison of prediction errors for TPPD claim

	Prediction model	MSE	RMSE	MAE	MAPE	Relative performance based on best MAPE
Claim frequency	Regression	25.69	5.07	16.6	0.8622	3.94
	BPNN	***369.5***	***19.22***	***10.7307***	***0.2191***	***1.00***
	ANFIS	10979.87	104.78	45.39	0.9269	4.23
Claim severity	Regression	1373.93	37.07	1197.62	0.6933	1.06
	BPNN	***2870793.9***	***1694.34***	***1105.2***	***0.6515***	***1.00***
	ANFIS	4366628.2	2089.65	1460.79	0.8331	1.28

Table 7 Comparison of prediction errors for OD claim

	Prediction model	MSE	RMSE	MAE	MAPE	Relative performance based on best MAPE
Claim frequency	Regression	265.5	16.29	3.53	0.3306	1.52
	BPNN	***3383.5615***	***58.1683***	***32.4928***	***0.2169***	***1.00***
	ANFIS	19915.27	141.12	61.77	1.1787	5.43
Claim severity	Regression	10532.17	102.63	79.1	0.4088	1.25
	BPNN	***8441272.93***	***2905.39***	***2097.01***	***0.3261***	***1.00***
	ANFIS	99893796.82	9994.69	7140.87	0.4417	1.35

Acknowledgement

This research is fully supported by FRGS grant, PY/2016/07210. The authors fully acknowledged Ministry of Higher Education (MOHE) and Universiti Teknologi Malaysia for the approved fund which makes this important research viable and effective.

References

[1] Azlan, M. Z., Habibollah, H. and Safian, S. 2010. Prediction Of Surface Roughness In The End Milling Machining Using Artificial Neural Network. *Expert Systems with Applications*. 37(2): 1755-1768.

[2] Bahia, H. S. I. 2013. Using Artificial Neural Network Modelling in Forecasting Revenue: Case Study in National Insurance Company/Iraq. *International Journal of Intelligence Science*. 3(3):136-143.

[3] Batty, M., Tripathi, A., Kroll, A., Peter Wu, C-S., Moore, D., Stehno, C., Lau, L., Guszcza, J. and Katcher, M. 2010. *Predictive Modelling for Life Insurance*. Deloitte Consulting LLP.

[4] Behbood, V. and Lu, J. 2011. Intelligent Financial Warning Model Using Fuzzy Neural Network and Case-Based Reasoning. *IEEE Symposium on Computational Intelligence for Financial Engineering and Economics*. 1-6.

[5] Berry-Stölzle, T. R., Koissi, M. C. and Shapiro, F. A. 2010. Detecting Fuzzy Relationships In Regression Models: The Case Of Insurer Solvency Surveillance In Germany. *Insurance: Mathematics and Economics*. 46(3): 554-567.

[6] Boyacioglu, M. A. and Avci, D. 2010. An Adaptive Network-Based Fuzzy Inference System (ANFIS) For The Prediction Of Stock Market Return: The Case Of The Istanbul Stock Exchange. *Expert Systems with Applications*. 37(12): 7908-7912.

[7] Christmann, A. 2004. An Approach To Model Complex High-Dimensional Insurance Data. *Allgemeines Statistisches Archiv*. 88: 375-397.

[8] Dalkilic, T. E., Tank, F. and Kula, K. S. 2009. Neural Networks Approach For Determining Total Claim Amounts In Insurance. *Insurance: Mathematics and Economics*. 45(2): 236-241.

[9] David, M. 2015. Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance*. 20: 147-156.

[10] Dugas, C., Chapados, N., Ducharme, R., Saint-Mleux, X. and Vincent, P. 2011. A High-Order Feature Synthesis And Selection Algorithm Applied To Insurance Risk Modelling. *International Journal of Business Intelligence and Data Mining*. 6(3): 237-258.

[11] Duma, M., B. Twala, and T. Marwala. 2011. Improving The Performance Of The Ripper In Insurance Risk Classification: A Comparitive Study Using Feature Selection. *8th International Conference on Informatics in Control, Automation and Robotics*. 203-10.

[12] Frees, E. W., Peng Shi, P. and Valdez, E. A. 2008. Actuarial Applications Of A Hierarchical Insurance Claims Model. *ASTIN Bulletin*. 39 (1): 165-197.

[13] Frees, E. 2014. *Frequency And Severity Models*. In E. Edward, G. Meyers, and R. A. Derrig (Eds.). Predictive Modelling Applications in Actuarial Science, Cambridge. Cambridge University Press.

[14] Gorr, W., Olligschlaeger, A. and Thompson, Y. 2003. Short-Term Forecasting Of Crime. *International Journal of Forecasting*. 19(4): 579-594.

[15] Homayouni, N. and Amiri, A. 2011. Stock Price Prediction Using A Fusion Model Of Wavelet, Fuzzy Logic And ANN. *International Conference on E-business, Management and Economics IPEDR, IACSIT Press, Singapore*. (25): 277-281.

[16] Huang, Q., Lin, L. and Sun, T. 2011. Some Actuarial Formula of Life Insurance for Fuzzy Markets. *Applied Mathematics.* 2(8): 1046-1050.

[17] Ibiwoye, A., Ajibola, O. O. E. and Sogunro, A. B. 2012. Artificial Neural Network Model for Predicting Insurance Insolvency. *International Journal of Management and Business Research*. 2(1): 59-68.

[18] Ismail, N. and Jemain, A. A. 2008. Construction Of An Insurance Scoring System Using Regression Models. *Sains Malaysia*. 37(4): 412-41.

[19] Kitchen, F. L. 2009. Financial Implications Of Artificial Neural Networks In Automobile Insurance Underwriting. *International Journal of Electronic Finance*. 3(3): 311-319.

[20] Laas, D., Schmeiser, H. and Wagner, J. 2016. Empirical Findings on Motor Insurance Pricing in Germany, Austria, and Switzerland. *Geneva Papers on Risk and Insurance - Issues and Practice*. 41(3): 398-431.

[21] Lennon, H. 2011. *Generalized Linear Models and their Extensions for Insurance Data*. Master of Degree, University of Manchester.

[22] Lin, C. 2009. Using Neural Networks As A Support Tool In The Decision Making For Insurance Industry. *Expert Systems with Applications*. 36(3): 6914-6917.

[23] Liew, T. and Chen, W. Y. 1998. Intelligence Detection Of Drill Wear. *Journal of Mechanical Systems and Signal Processing*. 12: 863-873.

[24] Mol, C. D., Giannone, D. and Reichlin, L. 2008. Forecasting Using A Large Number Of Predictors: Is Bayesian Shrinkage A Valid Alternative To Principal Components? *Journal of Econometrics*. 146(2): 318-328.

[25] Mohamed, M. A., Ismail, H., Razali, A. M., Ismail, N. and Ganiyat, A. U. 2011. Own Damage, Third Party Property Damage Claims And Malaysian Motor Insurance: An Empirical Examination. *Australian Journal of Basic and Applied Sciences*. 5(7): 1190-1198.

[26] Morata, L. B. 2009. A Priori Ratemaking Using Bivariate Poisson Regression Models. *Insurance: Mathematics and Economics*. 44(1): 135-141.

[27] Pinquet, J. 2012. *Experience Rating In Non-Life Insurance*. Working Papers hal-00677100, HAL.

[28] Salcedo-Sanz, S., Cuadra, L., Portilla-Figueras, J. A., Jiménez-Fernández, S. and Alexandre-Cortizo, E. 2012. A *Review Of Computational Intelligence Algorithms In Insurance Applications*. Statistical And Soft Computing Approaches In Insurance Problems. Nova Science Publishers. 1-50.

[29] Shapiro, F. A. 2007. An Overview Of Insurance Uses Of Fuzzy Logic. *Computational Intelligence in Economics* and Finance. (II): 25-61.

[30] Shi, P., Feng, X. and Ivantsova, A. 2015. Dependent Frequency–Severity Modelling Of Insurance Claims. *Insurance: Mathematics and Economics*. 64: 417-428.

[31] Takagi, T. and Sugeno, M. 1983. Derivation Of Fuzzy Control Rules From Human Operator's Control Action. *Proceedings of the IFAC Symposium on Fuzzy Information, Knowledge Representation and Decision Analysis*. 55-60.

[32] Wang, F. K.,Chang,K. K.and Tzeng, C. W. 2011. Using Adaptive Network-Based Fuzzy Inference System To Forecast Automobile Sales. *Expert Systems with Applications.* 38(8): 10587-10593.

[33] Yeo, A. C., Smith, K. A., Willis, R. J., and Brooks, M. 2003. A Comparison Of Soft Computing And Traditional Approaches For Risk Classification And Claim Cost Prediction In The Automobile Insurance Industry. *Soft Computing in Measurement and Information Acquisition, Studies in Fuzziness and Soft Computing*. Springer-Verlag, 127: 249-261.

[34] Yildiz, B. and Akkoc, S. 2010. Bankruptcy Prediction Using Neuro Fuzzy: An Application in Turkish Bank. *International Research Journal of Finance and Economics*. 60: 114-126.

[35] Zanganeh, T., Rabiee, M. and Zarei, M. 2011. Applying Adaptive Neuro-Fuzzy Model For Bankruptcy Prediction. *International Journal of Computer Application.* 20(3): 15-21

[36] Zhang, G., Patuwo, B. E. and Hu, M. Y. 1998. Forecasting With Artificial Neural Networks: The State Of The Art. *International Journal of Forecasting*. 14(1): 35-62.