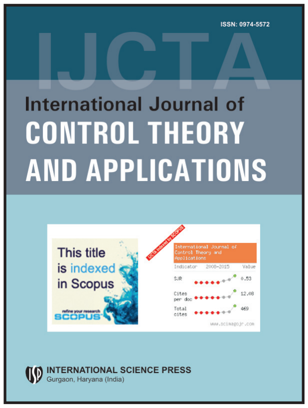

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 43 • 2016

Insurance Policy Advisor Using Fuzzy Decision Tree Technique

Sofianita Mutalib[a], Sufiah Mohd Fadzil[b], Norzaidah Mohd Nor[c] and Shuzlina Abdul Rahman[d]

[a,d]Research Interest Group of Intelligent Systems, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Darul Ehsan, Malaysia. Email: [a]sofi@tmsk.uitm.edu.my; [d]shuzlina@tmsk.uitm.edu.my
[b,c]Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Darul Ehsan, Malaysia. Email: [b]sufiah_takeisu@yahoo.com; [c]norzaidah@tmsk.uitm.edu.my

Abstract: People now realized how crucial the insurance is for their protection against irregular and uncertainty situations that will affect life and assets. Customers commit to insurances with various needs based on different backgrounds, economic conditions and risk preference degrees. However, determining the suitable insurance, would be troublesome. For that reason, a specific analysis should be done in matching the protection needs that would be provided by the right insurance to the customers. In understanding the requirements by customers, we explored the patterns in existing records. We have done a thorough experiment using Decision Tree to determine a suitable insurance policy based on the given information from the user. Eight variables were identified from literatures as the inputs, which are; sex, age, body mass index, occupation, annual income, smoking, drinking and contribution per month. These experiments managed to produce meaningful rules in determining four insurance types: life, health, annuity and investment-oriented. The combination of Decision Tree and fuzzy inference methods did improve the accuracy of selection of insurance type for the customers. These combination methods were applied in our model which is applicable for Malaysian population in general. This application would be beneficial to customers to manage and protect their assets in a more proper way.

Keywords: Advisor, Fuzzy Inference, Decision Tree, Insurance.

1. INTRODUCTION

The decision in determining suitable insurance policy to the customers could be a challenging for a consultant by considering various situations that involved uncertainty and vague variables. Customers also might be uncertain with the terms and conditions by different coverage. This is because of the customers' status are different in various backgrounds, economic conditions, risk preference degrees and needs and always have limited knowledge about suitable insurance products. The insurance consultants normally apply their experiences or "rules of thumb" based on certain customers' information, but do not consider it in more detail that might be resulting to ineligible claim in the future. In order to ease the process, a support system, such as advisory system would benefit people. Rokach et. al., have used multi agent system for recommending insurance policy rider to clients in call centres of Israeli finance company [1]. The recommendation depends on the behaviour patterns of similar users to the current active user and able to predict future purchase. Their experience showed that simple item-to-item collaborative filtering approaches provide an impressive ability to predict additional riders.

The major objective of our research is to identify the factors or set of variables relate to insurance and also to find significant knowledge by applying Decision Tree that able to construct rules in making suggestion. Decision Tree has been applied in many researches to classify and to make prediction with a high accuracy rate [2-6]. In section II, we further describe the related studies for this research. In section III and IV, the experimental setup and the result are presented. The discussion of the result was elaborated in section V and finally, we conclude and summarize this study in section VI.

2. RELATED STUDIES

There are a number of researches related to insurance as it is common that to protect themselves and the love one. Ibiwoye, Ajibola, and Sogunro trained an insolvency prediction model based on Artificial Neural Network (ANN) approach to evaluate the financial capability of insurance companies in Nigeria [7]. Forty percent from 118 insurance companies registered survived the recapitalization exercise. The contribution of the study was an interactive ANN simulation modelling that provides an early warning signal about distressed insurance companies. ANN was also used by Lin et. al., in their research by developing decision model for five insurances policies using back-propagation algorithm [8]. As a start for establishing the decision model, 300 customers from an insurance company in Taiwan were collected. They set up experiments in two phases: with six features as inputs and reduced features by factor analysis method. Their result showed that it is better to use more input features to determine the right insurance policy.

Another research was conducted by Bakar et. al., [9], that employed an associative classification model to develop a knowledge model for determining the best solution for insurance policy dataset. They enhanced the classification by introducing a heuristic in processing the correctly classified rules and the verified uncertain classified rules. The fundamental of the study is the insurance company can offer more than one different plan for a policy type or a clustered insurance policy. Huang, Lin, and Lin conducted a study on evaluation model for purchasing life insurance and health insurance by using fuzzy logic [10]. In the study, they used four factors as input of the proposed model including age, annual income, educational level, and assessment risk. They also used Analytical Hierarchical Processing (AHP) to generate the weights for the evaluation model. Other than that, Razak, Tan, and Lim [11] also have used decision model using fuzzy inference system (FIS) for insurance advisors to identify and to suggest appropriate policies to potential or existing clients by using five feature as the inputs. They also engaged an expert to verify the model.

A. Determine Risk Assessment

Risk assessments serve a basis for the determination of insurance premiums billed to the clients. Carreno and Jani [12] conducted a study using fuzzy expert system approach to develop a decision aid for evaluating risk for life insurance. They used FuzzyCLIPS tool to build a knowledge base with ten different inputs. The base inputs are age, weight and height, while the incremental inputs are clients habit, are exercising, intake of dairy products, red meat, vegetables, fat or sweet, smoking and drinking. The result showed more predictable and smoother behaviour between the variables by using fuzzy expert system. Besides, they proved that a better evaluation of the risk at any time based on particular habits of the individuals.

A research paper by Pokoradi [13] mentioned about risk as a measure of harm or loss allied with the human activity and he applied fuzzy logic-based to assess the risk assessments. The research is based on air unit commander with four different severity categories which is catastrophic, critical, moderate and negligible. He claimed that the commander can use the result to accept the risk at the next step of risk management process. Scollo and Pearce [14] also studied about factors that can affect to the increase mortality risk and malignancy risk between smoking status and body mass index (BMI). They claimed that a combination of current or recent smoking with high BMI is mostly related to the high mortality risk which is 5 to 8 times that a never smokers within a normal BMI. Based on certain experiments, they proved that obesity which mostly related to BMI. Furthermore, smoking is correlated to increase risk of malignancy and mortality. On top of that, an adaptive Neurofuzzy system was proposed by Bhardwaj, Singhal, and Gupta to determine asthma risk by using combination attributes of age, gender, economic status, tobacco and also smoke consumption [15].

B. Smoking and Drinking Status

Most of insurance companies need the information on smoking and drinking status. Barman and Choudhury applied fuzzy rule base system for diagnosis of heart disease in which years of smoking become one of the inputs for the system [16]. In the research, they classified the number of years of smoking into two categories which are low possibility for years below than 30 and high possibility for years more than 30. Meanwhile, Srivastava and Manan [17] categorized the smoker in their research into five different categories in term of cigar per week which are very low for consumed two to six cigars per week, low for consumed four to twelve cigars per week, moderate for consumed ten to sixteen cigars per week, high for consumed fourteen to twenty-two cigars per week and very high for consumed more than twenty cigar pers week. The research also categorized consumption of alcohol (per drink) into five categories. Next, Marston et. al., investigated smoking status recorded in UK primary care [18]. They classified smoking status into three categories which are smoker, ex-smoker and non-smoker. They claimed that an individual was not recorded as ex-smoker if they already quitted more than twenty-two years. Non-smoker can become an ex-smoker if they were still in age-specifies of time quitting smoking. They also mentioned those individuals were categorized as ex-smokers even they smoked for a short period of time. Other than that, fuzzy rule base is used to identify tuberculosis and smoking status is one of the input for the system [19, 20]. They categorized smokers into three categories which are not smoker who not consumed any cigar per day, little who consumed six cigars per day and moderate for those who consumed six to ten cigars per day. Following the above point of views, risk assessment, smoking habit and drinking habit are the variables considered in our research.

3. DATA DESCRIPTION

Our research is focusing to four types of insurance policy for customers which are life insurance, annuity insurance, health insurance, and investment-oriented insurance. The data studied in this research was insurance clients dataset obtained from an insurance company in Malaysia. It contains 160 records with 12 attributes of customer data including proposal number, gender, age, height, weight, occupation, annual income, smoking duration, number of smoke consume per day, drinking status, contribution premium per month, insurance package. The client number and proposal's number were removed as these attributes were not necessary in the mining process. The *'age'* attribute was transformed to categorical with reference to literatures and expert, which ranging into three categories, as shown in Table 1.

Table 1
Data Transformation: Age Attribute

Age Category	*Frequency*	*Range*
Young	Low	4 weeks – 18 years
Adult	Medium	19 years – 25 years
Elder	High	26 years – 66 years

Meanwhile, occupation level was transformed to categorical with four occupation classes in Table 2. *'Annual income'* attribute was transformed to categorical also, based on *worldbank.org*, into four income classes in Malaysia, as shown in Table 3 [21]. The '*height*' and '*weight*' attributes were used to calculate Body Mass Index (BMI) attribute. BMI is calculated by dividing the weight (in kilogram unit) with squared height (in meter unit).

Table 2
Data Transformation: Occupation Class

Occupation Class	*Occupation Category*	*Explanation*	*Example*
Class 1	White Collar	Workers are in harmless industry and limited within the office premises	Administrative, Student, Salesman, Teacher
Class 2	Overseers	People is working in closed buildings, occasionally engaging in manual labor or outdoor activities	Nurse, Steward, Sales Technical, Technician
Class 3	Skillful Workers	People is working light manual labor with/without the use of tools or machinery	Chef, Guard, Operator, Security
Class 4	Industrial workers/ blue collar	People who deals with heavy & dangerous machinery	Military, Police

Table 3
Data Transformation: Annual Income Attribute

Income Category	*Range*
Poor	X <= RM10,800
Vulnerable	RM10,800 > x <= RM25,440
Aspirational	RM25,440 > x <= RM70,800
Middle / Upper	x > RM70,800

The above discussed variables; age, occupation and income, were considered as the input of descriptive variables. We mapped the insurance package name, as output variable, that are available in one of the top insurance companies. We present the insurance packages into four major categories, as shown in Table 4.

Table 4
Insurance Packages

	Life Insurance	*Health Insurance*	*Annuity Insurance*	*Investment-oriented insurance*
My Child	/			/
Life Ready	/	/	/	
Link One		/	/	/
Lady	/	/		
Cash				/
Firstlink	/			/
Takafullink		/		/
Smartlink	/			/
Ummahlink				/

Finally, we concluded the variables used for our experiments are *Sex, Age, Occupation Level (OL), Annual Income Class (AI Class), Drinking, Cig per day, Year Of Smoking, Contribution Per Month (CPM), BMI,* and the output are *life insurance, health insurance, annuity insurance, and investment-oriented insurance.*

4. EXPERIMENT AND RESULTS

These experiments were mainly divided into two phases: Phase 1 applied Decision Tree (DT) J48 and Phase 2 applied Fuzzy Decision Tree (FDT). The experiments were performed using a data mining tool named Weka 3.7. The classification for Decision Tree J48 was made with cross validation testing of 10 folds. In the second phase, we implemented an inference engine to the experiments using MATLAB to conclude risk preference. The variables with fuzzy sets are *BMI, smoking duration, consumption* and *risk,* as shown in Table 5. Fuzzy rules for risk preference were gained from the domain expert of insurance company and also a pharmacist who involves in that area for almost 22 years.

Table 5
Variables used in Fuzzy Inference System

Variable	*Fuzzy Sets*	*Value*
BMI	Under	[0,23]
	Ideal	[17,31]
	Over	[25,33]
	Obese	[30,60]
Smoking Duration	Low Possibility	[0,30]
	High Possibility	[30,60]
Smoking Consumption	NonSmoker	[0,1.5]
	Less than 8 years	[1.5,5.25]
	Extreme	[8.122,17.5]
Risk	Very Low	[0,0.1286]
	Low	[0.0873,0.3603]
	Med	[0.262,0.6683]
	High	[0.538,0.8937]
	Very High	[0.7667,1.2]

A. Phase 1: Rule Mining using Decision Tree J48

In these phase, we would like to find the rules generated by DT for each insurance package. In Phase 1, the dataset was split into training and testing sets with 6:4, 7:3 and 8:2 ratio. The following Table 6 shows the accuracy results with Life Insurance gained highest accuracy at 89%, Health Insurance is at 76%, Annuity Insurance is at 81% and lastly Investment insurance is at 79%.

Table 6
Classification using Decision Tree

Type of insurance	*Ratio*	*Mean Absolute Error*	*Root Mean Square Error*	*Relative Absolute Error (%)*	*Root Relative Square Error (%)*	*Correctly Classified Instance (%)*	*Incorrectly Classified Instance (%)*
Life	8:2	0.102	0.274	54.705	89.792	59.375	40.625
	6:4	0.136	0.299	27.708	60.327	89.583	10.417
	7:3	0.172	0.347	34.729	69.730	86.607	13.393
Health	8:2	0.085	0.236	45.521	77.218	65.625	34.375
	6:4	0.312	0.469	64.525	95.484	73.958	26.042
	7:3	0.253	0.407	52.609	82.966	76.786	23.214

Type of insurance	*Ratio*	*Mean Absolute Error*	*Root Mean Square Error*	*Relative Absolute Error (%)*	*Root Relative Square Error (%)*	*Correctly Classified Instance (%)*	*Incorrectly Classified Instance (%)*
Annuity	8:2	0.056	0.195	29.737	63.881	81.250	18.750
	6:4	0.381	0.490	82.949	102.443	65.625	34.375
	7:3	0.402	0.512	87.434	106.748	61.607	38.393
Investment	8:2	0.089	0.243	47.504	79.777	65.625	34.375
	6:4	0.202	0.387	58.383	93.447	79.167	20.833
	7:3	0.246	0.447	74.630	110.502	77.679	22.321

B. Phase 2: Rule Mining using J48 with Fuzzy Variables

In phase 2, dataset is divided to training and testing data with ratio of 8:2 and we applied FDT J48. As shown at Table 7, the accuracy of classification is more than 60 percent.

Table 7
Classification using Decision Tree with Fuzzy Variables

Insurance	*Mean Absolute Error*	*Root Mean Square Error*	*Relative Absolute Error (%)*	*Root Relative Square Error (%)*	*Correctly Classified Instance (%)*	*Incorrectly Classified Instance (%)*
Life	0.200	0.360	40.252	72.258	82.813	17.188
Health	0.253	0.404	53.937	83.428	76.563	23.438
Annuity	0.394	0.536	83.329	110.234	62.500	37.500
Investment	0.197	0.345	58.497	84.507	85.156	14.844

5. DISCUSSION

Table 8 compares two methods applied, and it shows that the highest accuracy using DT by 89.58% for life insurance, while investment insurance gained higher accuracy using FDT method by 85.16%. In general, DT produced higher accuracy compare to FDT. Though the results show DT method gave better accuracy, the rules constructed using second method are more relevant. Table 9 shows sample of rules extracted for life insurance and health from both methods. For health and annuity, FDT produced more rules compared to DT.

Table 8
Summarization of two phases

Insurance Type	*Decision Tree Accuracy*	*Decision Tree And Fuzzy Variables Accuracy*	*No. of Rules From Decision Tree*	*No. of Rules From Decision Tree with Fuzzy Variables*
Life	89.58	82.81	4	6
Health	76.79	76.56	5	9
Annuity	81.25	62.50	13	22
Investment	84.38	85.16	12	8

From the rules found, we can conclude that, variables*; age, and contribution premium per month, occupation level and sex* are the important variables to consider in purchasing life insurance policy. As for health insurance policy, variables of *age, annual income, occupation level and risk* are commonly useful. Contribution per month appeared most of the rules in determining annuity insurance policy as the policy is more about to protect someone's finance. The rule constructed for investment-oriented insurance mostly used variables are *age, contribution premium per month, annual income and risk preference*. Variable risk preference was found during the generation of rule for Health, Annuity, and Investment insurance, but not for Life insurance. It can be concluded that the risk preference is significant to the insurance consultant for decision making of the suitable type of insurance.

Table 9
List of the rules extracted for life insurance for both phases

Insurance	*Rules Extracted From Decision Tree*	*Rules Extracted From Decision Tree With Fuzzy Variables*
Life	If Age=high && CPM<=126.31 Then yes; If Age=high && CPM>126.31 Then no; If Age=med Then yes; If Age=low Then no;	If Age=high Then no; If Age=med && CPM<=150.22 Then yes; If Age=med && OL>2 Then yes; If Age=med && OL<=2 && Sex>0 Then yes; If Age=med && OL<=2 && Sex<=0 Then no; If Age=low Then yes;
Health	If Age=high Then yes; If Age=med && Smoking=no && AI<=24000 && CPM<=220 Then yes; If Age=med && Smoking=no && AI<=24000 && CPM>220 Then no; If Age=med && Smoking=yes Then no; If Age=low Then no;	If Age=high Then yes; If Age=med && AI Class=aspirational && OL>1 Then yes; If Age=med && AI Class = aspirational && OL<=1 Then no; If Age=med && AI Class = vulnerable && OL<=2 Then yes; If Age=med && AI Class = vulnerable && OL>2 Then no; If Age=med && AI Class = middle/upper Then yes; If Age=med && AI Class=poor && Risk<=0.1015 Then yes; If Age=med && AI Class=poor && Risk>0.1015 Then no; If Age=low Then no;

6. CONCLUSION

This particular research was done with fuzzy variables and also Decision Tree, J48 for classification of insurance datasets and producing rules for decision making. Twelve variables were chosen as the inputs for the insurance policy advisor. Decision tree gained higher accuracy rate and fuzzy decision tree produced more significant rules. It is hoped that this research will provide significant knowledge in this field and the future enhancements can be made to make this research become more successful. This research has a commercial potential to any insurance company that want to enter digital marketing in order to reach their end user faster and wider. The customers can contact the virtual insurance advisor at anytime and anywhere rather than following the working office hours. Mobile application can also be developed based on this research.

Acknowledgment

The authors would like to thank Research Management Centre of Universiti Teknologi MARA and REI 19/2015 in supporting the publication of this research.

REFERENCES

[1] L. Rokach, G. Shani, B. Shapira, E. Chapnik, and G. Siboni, *Recommending Insurance Riders,* pp. 253-260, 2013.

[2] M. N. a. M. A. H. Amin, "American Journal of Engineering Research (AJER) " *Comparison of Different Classification Techniques Using WEKA for Hematological Data.,* Vol. 4, pp. 55-61, 2012.

[3] R. a. S. Arora, "International Journal of Computer Applications " *Comparative Analysis of Classification Algorithms on Different Datasets using WEKA.,* Vol. 54, pp. 21-25, 2012.

[4] S. Mutalib, et. al., *An exploratory study in classification methods for patients' dataset. in 2009 2nd Conference on Data Mining and Optimization.,* 2009.

[5] P. a. K. K. V. Sewaiwar, "International Journal of Emerging Research in Management &Technology," *Comparative Study of Various Decision Tree Classification Algorithm Using WEKA. ,* Vol. 4, pp. 87-91, 2015.

[6] E. a. T. V. Venkatesan, "Indian Journal of Science and Technology " *Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification.,* Vol. 8, pp. 1-8, 2015.

[7] A. Ibiwoye, O. O. E. Ajibola, and A. B. Sogunro, *Artificial Neural Network Model for Predicting Insurance Insolvency,* Vol. 1, pp. 59-68, Winter 2012 2012.

[8] Y.-J. Lin, Chin-Sheng-Huang, and C.-C. Lin, *Determination of Insurance Policy Using Neural Networks and Simplified Models with Factor Analysis Technique,* Vol. 5, p. 10, October 2008 2008.

[9] A. A. Bakar, Z. Othman, M. S. N. M. Yusoff, and R. Ismail, *Development of Knowledge Model for Insurance Product Decision using the Associative Classification Approach,* pp. 1481-1486, 2010.

[10] C.-S. Huang, Y.-J. Lin, and C.-C. Lin, *An Evaluation Model for Determining Insurance Policy Using AHP and Fuzzy Logic: Case Studies of Life and Annuity Insurances,* 2007.

[11] S. F. A. Razak, S. C. Tan, and W.-S. Lim, *A Decision Model for Insurance Advisors: A Case Study* 2013.

[12] L. A. Carreno and D. Y. Jani, *A Fuzzy Expert System Approach to Insurance Risk Assesment Using Fuzzyclips,* 1993.

[13] L. Pokoradi, "Academic Applied Research in Military and Public Management Science," *Fuzzy logic-based risk assessment,* Vol. 1, p. 11, 2002.

[14] M. Scollo and M. Pearce, "C. Council, Editor 2012, Melbourne: Cancer Council Victoria: Australia.," *The health effects of active smoking, Tobacco in Australia: Facts and Issues. Fourth Edition,* 2012.

[15] S. Bhardwaj, N. Singhal, and N. Gupta, "International Journal of Research in Engineering and Technology," *Adaptive Neurofuzzy System For Asthama* Vol. 3, p. 6, 2014.

[16] M. Barman and J. P. Choudhury., " International Journal of Computer Applications, ," *A Fuzzy Rule Base System for the Diagnosis of Heart Disease,* Vol. 57, p. 7, 2012.

[17] P. Srivastava, A. Srivastava, and A. Manan, "International Journal of Computer Applications," *A Soft Computing Decision Support System in the Diagnosis of Breast Cancer,* Vol. 72, pp. 19-27, 2013.

[18] L. Marston, J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth, Ian R White*, et. al.*, *Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study,* 2015.

[19] S. K. Tiwari, N. Walia, and R. Malhotra, *Design and Identification of Tuberculosis using Fuzzy Based Decision Support System,* 2015.

[20] N. Walia, H. Singh, S. K. Tiwari, and A. Sharma, "International Journal on Soft Computing," *A Decision Support System For Tuberculosis Diagnosability,* Vol. 6, pp. 1-14, 2015.

[21] worldbank.org. (2014). Available: http://www.worldbank.org/en/country/malaysia