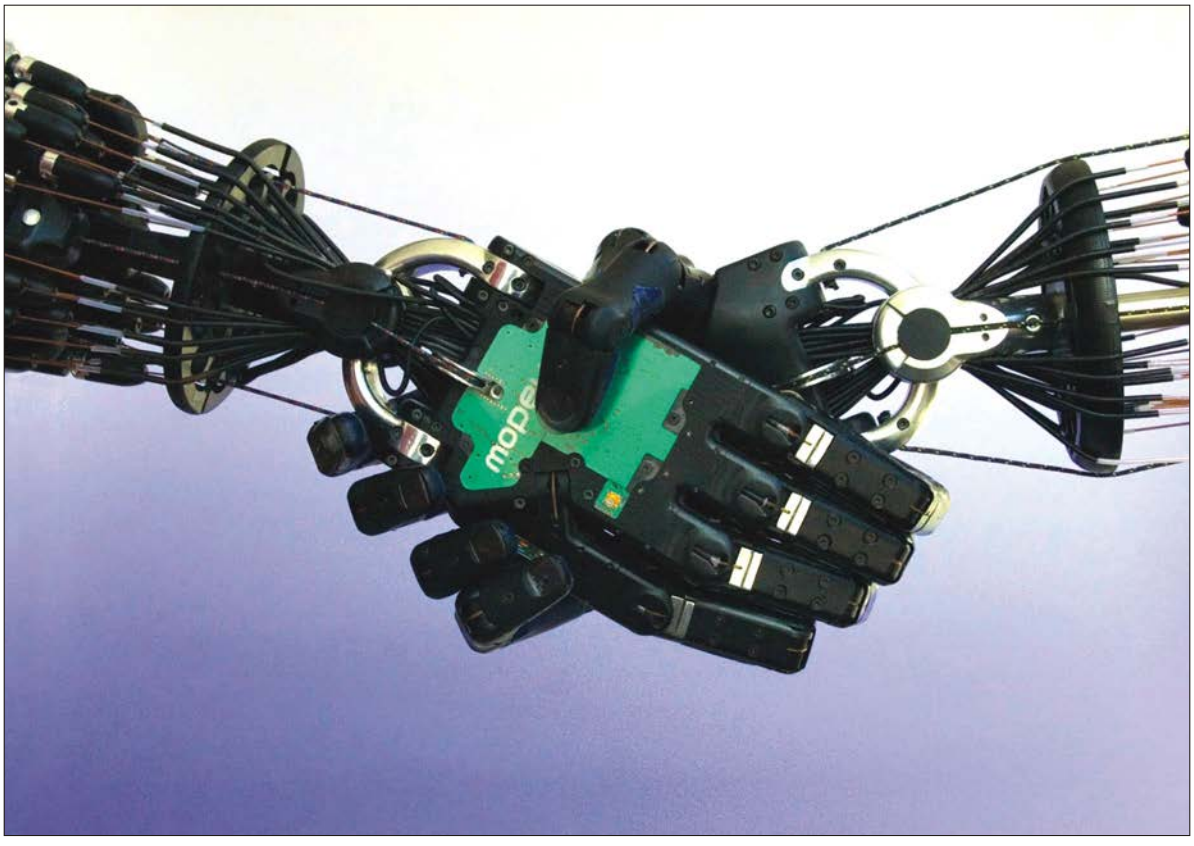

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2019.18.23

Research Article
Risk Assessment with Decision Tree in Professional Liability Insurance: In Accounting

[1]Murat Sari, [2]Eyyup Gulbandilar and [3]Nilüfer Dalkilic

[1]Department of Mathematics, Faculty of Art and Science, Yildiz Technical University, Davutpasa, 34220 Istanbul, Turkey
[2]Department of Computer Engineering, Faculty of Engineering and Architecture, Eskisehir Osmangazi University, 26480 Eskisehir, Turkey
[3]School of Applied Sciences, Dumlupinar University, 43100 Kutahya, Turkey

Abstract

Background and Objective: Evaluation of new tools to assess the risk of professional liability insurance is needed in daily life shaped by many parameters. The pragmatic aim of this study is to deal with the assessment of insurance risks in the professional liability insurance through decision tree algorithm and the entropy. Thus the present study is to provide effective decision-making based on risk factors of insurance in various branches of professional liability. **Materials and Methods:** To achieve this study, an algorithm was produced by taking into consideration a quite big number of variables. This algorithm was based on a decision tree and entropy. To produce this algorithm, 258 policies (exam group) were tested on the 54 policies (testing group). **Results:** The computed results were seen to be in very good agreement with the policies. Over 87% of the results are in agreement with the policies. Our tool is designed to be used by professional liability insurance companies at minimum risk and to be used at optimum prices of clients. **Conclusion:** This study is the first and important attempt to assess the level of risk for a wide range of insurance companies and to find the optimal price for many clients.

Key words: Insurance, entropy, data mining, decision tree, professional liability insurance, C4.5 algorithm

Citation: Murat Sari, Eyyup Gulbandilar and Nilüfer Dalkilic, 2019. Risk assessment with decision tree in professional liability insurance: In accounting. J. Artif. Intel., 12: 18-23.

Corresponding Author: Eyyup Gulbandilar, Department of Computer Engineering, Faculty of Engineering and Architecture, Eskisehir Osmangazi University, 26480 Eskisehir, Turkey Tel: +90 222 239 37 50

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

The insurance covers a large area. Each branch of professional liability insurance is comprehensive and requires expertise. The scope of this study consists of professional liability insurance in accounting and legal risks of this insurance coverage. Determination of risk and premium in insurance is of great importance. There are many statistical methods in the development of insurance risk assessment models.

Insurance claims were defined in terms of entropy of probability distribution for losses. It was tried to find out whether there is a relationship between the level of loss and purchase insurance requirements[1]. The entropy approach was applied to crop insurance by Najafabadi *et al.*[2] to estimate loss sizes. In the insurance sector, degree of competition, market structure and market power were analysed by the entropy method[3,4]. They paid attention to the computation of credibility parameters based on the concept of relative entropy between demand sizes of the entire portfolio. Interested readers are referred to some studies in the field of insurance with the entropy method[5-7].

According to the authors' best knowledge, there is no study on decision tree algorithm and entropy regarding the assessment of insurance risks in professional liability insurance. This research is expected to provide an effective decision-making based on insurance risk factors in professional responsibility branches such as independent accountant, independent financial advisor and independent accountant financial advisor. Thus, insurance companies will be able to understand risk policies and evaluate risk-based premiums. Insurance companies can claim low premiums for high risk and high risk policies for low risk policies. This study will also help insurers decide whether policies are renewable. In the light of previously explained necessities, this paper aims at evaluating risks in the professional liability insurance with decision tree algorithm and entropy. Thus, the present paper is believed to provide an effective decision-making tool based on the risk factors of insurance in various branches of professional responsibility.

Since 1970s machine learning has been paid attention, specifically a decision-tree procedure, ID3 (Iterative Dichotomiser), was developed[8]. This study was extended from previous studies on concept learning systems[8,9] and C4.5 was then produced, this became a benchmark against which new controlled learning algorithms were compared. Breiman *et al.*[10] published the book Classification and Regression Trees (CART), which explains the formation of binary decision trees. Two similar approaches for learning decision trees, ID3 and CART, were invented independently of each other. The ID3, C4.5 and CART adopted a non-feedback approach in which decision trees are constructed in the form of a top-down recursive partitioning-conquer way. Many approaches to starting the decision tree continue in such a top-down manner, starting with the class tags and their incorporated tags.

The ID3 utilizes information gain as a feature selection measure. This measure is based on the remarkable work of Shannon and Weaver[11] on the theory of knowledge that investigates the value. Let node N indicate the tuples of partition D. The attribute with the highest information gain is selected as the section attribute for the node N. This feature reduces the information needed to categorize screens in the result sections and reflects the least randomness in those sections. Such an approach minimizes the number of expected tests needed to categorize a given tuple and allows a simple tree to be found. The expected information required to categorize a tuple in D is given by:

$$\text{Info}\ (D) = -\sum_{i=1}^{m} p_i \log 2(p_i) \quad (1)$$

where, p_i is the probability of an arbitrary tuple in D belongs to class C_i and is estimated by $\frac{|C_i, D|}{|D|}$. Info (D) stands for the average amount of information required to define the class label of a tuple in D. Notice that information is based only on the proportions of tuples of each class, also known as the entropy of D. Entropy is one of the most common discretization measures used in data based applications. It was first produced by Shannon and Weaver[11] in the pioneering study on the concept of information theory. Entropy-based discretization is a supervised, top-down partition process. It investigates the class distribution information in the computation of the split points. For discretization of a numerical attribute, A, the method takes the value of A with minimum entropy as a split-point and repeatedly divides the resulting ranges to achieve a hierarchical separation. Such discretization creates a concept hierarchy for the attribute A.

Now let us divide the tuples in D on some attribute A with v different values, $\{a_1, a_2, ..., a_v\}$ as realized from the training data. If A has a discrete value, these values correspond directly to the v results of a test on A. The attribute A can be used to divide D into v sections, $\{D_1, D_2, ..., D_v\}$, where D_j involves those tuples in D that have outcome a_j of A. These sections correspond to the branches grown from node N. Let this segmentation produce a complete classification of the

tuples. However, the partitions are quite likely to be impure. To obtain a complete classification, the amount of information is calculated by:

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{x} \text{Info}(D_j) \quad (2)$$

The expression $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. $\text{Info}_A(D)$ is the expected information to classify a tuple from D based on the partitioning by A. The smaller the expected information required, the greater the purity of the partitions. Information gain is given by:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad (3)$$

The attribute A with the highest information gain, Gain (A), is chosen as the splitting attribute at node N. For more details, readers interested in the entropy technique are referred to the work of Han and Kamber[12].

MATERIAL METHODS AND STUDY DESIGN

The main subject of this study is the legal risk analysis of professional liability insurances. The data in the proposal form used belongs to the firm of the members of the accounting profession. The risk was dealt with through the entropy-decision tree. The input variables used in the professional liability insurance located in the offer forms are as follows:

- Profession (Independent accountant, independent financial advisor, independent accountant-financial advisor)
- Proportions of corporate income taxes
- Proportions of individual income taxes
- Giro in the financial year-end
- Giro in the current financial year
- Insured person working alone? (Yes/No)
- Insurance application cancelled? (Yes/No)
- Insurance demand cancelled? (Yes/No)
- Insurance premium (Turkish Lira)
- Amount of damage (Turkish Lira)

The output variable is the legal risk. The insurance companies covered in this study consist of 312 policies (258 policies: Examination group and 54 policies: testing group) for the evaluation of the decision tree in the entropy method.

Decision tree introduced by using entropy approach: In this approach, in order to determine the risk in professional liability, an algorithm was determined by classifying the input parameters in terms of the produced decision tree. Since some of the parameters contained quantitative values, the C4.5 algorithm was preferred. The median of the input parameters consisting of quantitative values was calculated. Therefore, the input parameters are basically classified into 2 groups: (i) The values of the input parameters are less than or equal to the median, (ii) The values of the input parameters are greater than the median. Taking into account the input parameters, the following can be listed as: profession, proportions of income, proportions of individual income taxes, giro in the financial year-end, giro in the current financial year, the insured person working alone, the cancellation of the insurance claim, the cancellation of the insurance application, the insurance premium and the amount of damage. The first and second groups are thought to be damaged and not-damaged, respectively. For risk in professional liability, the classes are $C_{\text{damaged}} = 79$ and $C_{\text{not-damaged}} = 179$. In this respect, the probabilities are found as $p_{\text{damaged}} = \frac{79}{258}$ and $p_{\text{not-damaged}} = \frac{179}{258}$. The entropy values, in the sense of the average amount of information, can be found using Eq. 1. Using Eq. 2, entropy values were calculated for each value of input parameters in the sense of expected information. Also, using Eq. 3, the information gain for the input variables can be seen in Table 1. As shown in Fig. 1, "the amount of damage" was seen as root of decision tree and it has the maximum value of the information gain.

The risk in professional liability is classified for the case of "less or equal to" of "the amount of damage" and using Eq. 1 the entropies of "the amount of damage" were calculated. For the case of "less or equal to" of the amount of damage, using Eq. 2, entropy values were calculated for each value of the input variables in the sense of expected information. For the same case, also, using Eq. 3, the information gain for the input parameters can be seen in Table 2. As seen from Fig. 1, the "insurance demand cancelled?" was seen to as the left hand of the decision tree and it has the maximum value of the information gain. As clearly seen from Fig. 1, the risk in professional liability was classified for the case of "less or equal to" of the "insurance demand cancelled?" and has been found to be "less or equal to". On the other hand, there exist 2 different cases for the case "greater" of the parameters of "damaged". The corresponding decision trees are seen in Fig.1. The risk in professional liability was classified for the case of "less or equal to" of " the amount of damage" and using Eq. 1 the entropies of "the amount of damage" were calculated. For the case of "yes" of the "insurance demand cancelled", using

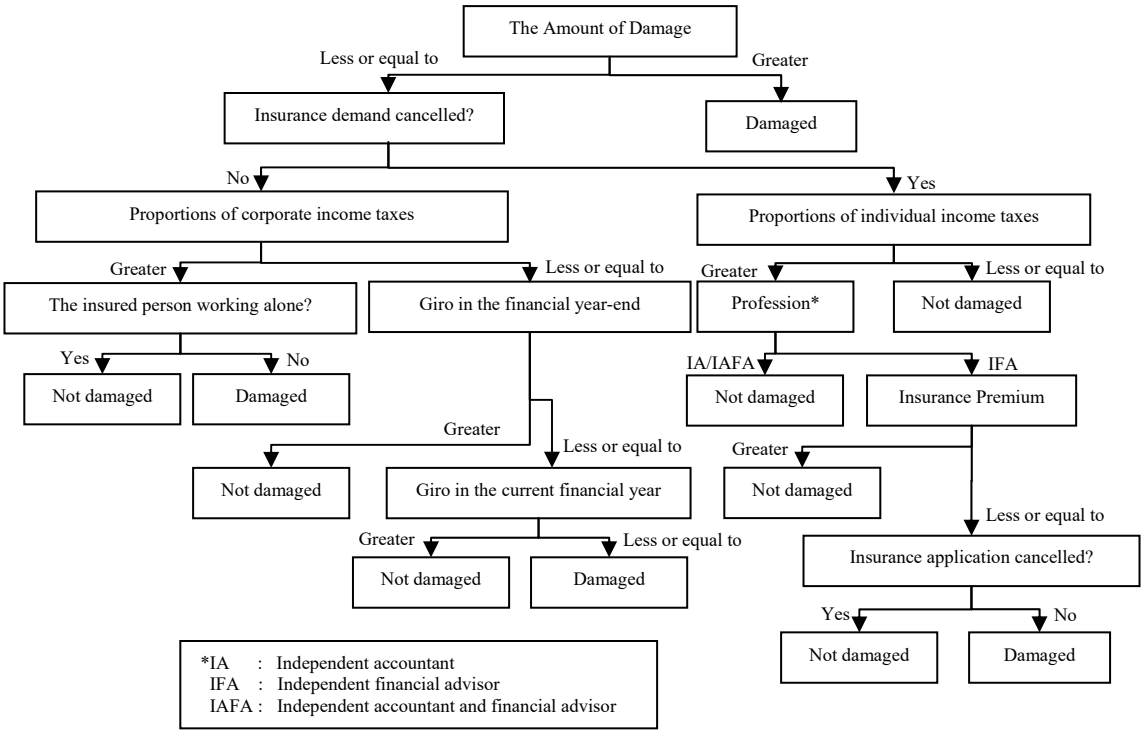

Fig. 1: Decision tree

Table 1: Calculation of gain to main root of the decision tree

Input parameters	Gain
Profession	0.0366
Proportions of corporate income taxes	0.0323
Proportions of individual income taxes	0.0145
Giro in the financial year-end	0.0222
Giro in the current financial year	0.0194
The insured person working alone?	0.0050
Insurance application cancelled?	0.0202
Insurance demand cancelled?	0.0320
Insurance premium	0.0144
The amount of damage	0.3197

Table 2: Calculation of gain to "the amount of damage"

Amount of damage	Less or equal to gain	Greater
Input parameters		
Profession	0.0190	Damaged
Proportions of corporate income taxes	0.0426	
Proportions of individual income taxes	0.0030	
Giro in the financial year-end	0.0104	
Giro in the current financial year	0.0040	
The insured person working alone?	0.0046	
Insurance application cancelled?	0.0107	
Insurance demand cancelled?	0.0447	
Insurance premium	0.0087	

Eq. 2, the entropy values were computed for each one of the input variables in the sense of expected information. For the same case, also, using Eq. 3, the information gains of the input parameters were presented in Table 3. As shown in Fig.1, the

Table 3: Calculation of gain to "insurance demand cancelled?"

	Insurance demand cancelled?	
	Gain	
Input parameters	Yes	No
Profession	0.0194	0.0124
Proportions of corporate income taxes	0.0096	0.0677
Proportions of individual income taxes	0.0359	0.0038
Giro in the financial year-end	0.0318	0.0053
Giro in the current financial year	0.0069	0.0242
The insured person working alone?	0.0300	0.0078
Insurance application cancelled?	0.0044	0.0042
Insurance premium	0.0250	0.0039

"insurance demand cancelled" was seen to be the left hand of the decision tree and it has the maximum value of the information gain. As seen in the corresponding figure, the "not-damaged" was seen to be the right hand of the decision tree. As seen from Fig. 1, the risk in professional liability was classified for the case "greater" of "profession" (Table 4). Also, the information gain of the input parameter (proportions of corporate income tax) was found to be maximum for the case "no" of "insurance demand cancelled". The corresponding decision trees were given in Fig. 1.

For the case "less or equal to" of "proportions of corporate income tax", the risk in professional liability was found to be "giro in the financial year-end". When taking the case "greater"

Table 4: Calculation of gain to "proportions of individual income taxes"

	Proportions of individual income taxes	
Input parameters	Less or equal to gain	Greater
Profession	Not damaged	0.0934
Giro in the financial year-end		0.0773
Giro in the current financial year		0.0311
The insured person working alone?		0.0635
Insurance application cancelled?		0.0224
Insurance premium		0.0635

Table 5: Calculation of gain to "proportions of corporate income taxes"

	Proportions of corporate income taxes	
Input parameters	Less or equal to gain	Greater
Giro in the financial year-end	0.0084	
Giro in the current financial year	0.0001	0.2730
The insured person working alone?	0.0000	0.3122
Insurance application cancelled?	0.0000	0.0506
Insurance premium	0.0077	0.0009

Table 6: Comparison of insurance policy with computer prediction

	Insurance policy		
Computer prediction	Damaged	Not damaged	Total
Damaged	16	3	19
Not damaged	4	31	35
Total	20	34	54

as presented in Fig. 1, the input variable "proportions of corporate income tax" was seen to be maximum gain of the information. The risk in professional liability of those input factor "the insured person working alone" was shown in Fig. 1 and Table 5. In a similar manner, the rest of the details of the decision tree can be read from Fig. 1. The produced program codes using the decision tree is as follows:

…

- If "The amount of damage" is "Greater" than "Damaged"
- If "The amount of damage" is "Less or equal to" and "Insurance demand cancelled?" is "Yes" and "Proportions of individual income taxes" is "Less or equal to" than "Not Damaged"

…

C # programming codes were generated by using decision tree to determine the effects of input variables. The produced codes were tested for 54 policies.

RESULTS AND DISCUSSIONS

The accuracy of the software developed using 258 policies (exam group) was tested on 54 policies (testing group). As seen in Table 6, 47 policies from the testing group were calculated correctly (87.04%). The results presented in the table also showed that 16 out of 20 damaged policies were estimated correctly (80.0%). The produced codes predicted 31 policies of 34 "not damaged" policies correctly (91.18%).

One of the most common approaches is the entropy encountered in applied sciences. The insurance demand in terms of the entropy of the probability distribution for losses was characterized in Nakata *et al.*[1]. They tried to find out if there was a relationship between the level of loss and the purchase insurance requirements. In order to estimate losses, the entropy approach was applied by Najafabadi *et al.*[2] to crop insurance. The degree of competition, market structure and market power in the insurance society were examined through the entropy by Bello *et al.*[3]. Fernández-Durán and Gregorio-Domínguez[4] concentrated on the calculation of credibility factors based on the concept of relative entropy between the claim size distributions of the policyholder. More details of the related studies carried out in the field of insurance through the entropy can be found in the references[5-7].

Some decision models for insurance were attempted to be created in Huang *et al.*[13] using 300 insurance companies from a Taiwanese insurance company. In this article, decision trees were used to make purchases. Five main types of insurance were included in this study, including life, annuity, health, accident and investment-oriented insurances.

A comparative analysis of predictive performance of a battery of data mining techniques using real-life automotive insurance fraud data was provided by Gepp *et al.*[14]. For a prediction, a successful comparison was made by between logistic regression, neural network and decision tree classifiers[15]. Yet, it was proposed to use a causal inference framework to measure the price elasticity of auto insurance. Their model allows one for estimating price-elasticity functions at the individual policyholder level[16].

CONCLUSION

The assessment of the risk levels of professional insurance companies and the best prices for customers are very important in daily life. The level of risk in professional liability insurance and optimal price for many clients were successfully discovered with the decision tree algorithm with entropy. It was concluded that effective decision-making based on the risk factors of insurance is provided in various branches of professional liability for multi-variables. The computed results were realized to be in agreement with the policies, over 87% of the results. According to the results produced by the tool designed in this study, there was a minimum risk for

professional liability insurance companies and at the same time the prices were the most favorable for clients. Note also that insurance companies are able to distinguish between risk policies and calculate premiums based on possible risk. Insurance companies can request low premiums for low risk policies and high premiums for high risk policies. This study is also believed to help insurance companies decide whether policies are renewable. This study is believed to help researchers who want to uncover critical areas of insurance companies and affordable prices for many customers.

REFERENCES

1. Nakata, H., Y. Sawada and M. Tanaka, 2010. Entropy characterisation of insurance demand: Theory and evidence. RIETI Discussion Paper Series 10-E-009, pp: 1-33.
2. Najafabadi, A.T.P., H. Hatami and M.O. Najafabadi, 2012. A maximum-entropy approach to the linear credibility formula. Insurance: Math. Econ., 51: 216-221.
3. Bello, H.M., M. Sepiriti and E.M. Letete, 2009. Competition, market structure and market power in the insurance industry in Lesotho. ICFAI J. Fin. Econ., 7: 7-21.
4. Fernández-Durán, J.J. and M.M. Gregorio-Domínguez, 2004. Relative entropy credibility theory. Proceedings of the 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings, Vol. 735, July 25-30, 2004, Garching, Germany, pp: 60-67.
5. Bagchi, D., 2009. Application of entropy to evaluate solvency of the insurance companies. IUP J. Risk Insurance, 6: 7-18.
6. Darooneh, A.H., 2004. Non-life insurance pricing: Multi-agent model. Eur. Phys. J. B: Condensed Matter Complex Syst., 42: 119-122.
7. Kumar, N., V. Verma and V. Saxena, 2013. Construction of decision tree for insurance policy system through entropy and gini index. Int. J. Comput. Applic., 70: 7-10.
8. Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA., USA.
9. Hunt, E., J. Martin and P. Stone, 1966. Experiments in Induction. Academic Press, New York.
10. Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone, 1984. Classification and Regression Trees. 1st Edn., Wadsworth Int. Group, Belmont, California, USA., ISBN: 978-0534980542, Pages: 368.
11. Shannon, C.E. and W. Weaver, 1949. The Mathematical Theory of Communication. 1st Edn., University of Illinois Press, Urbana, IL., ISBN-10: 0252725484.
12. Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publisher, San Fransisco, USA., ISBN-13: 978-1558609013, Pages: 800.
13. Huang, C.S., Y.J. Lin and C.C. Lin, 2008. Implementation of classifiers for choosing insurance policy using decision trees: A case study. WSEAS Trans. Comput., 7: 1679-1689.
14. Gepp, A., J.H. Wilson, K. Kumar and S. Bhattacharya, 2012. A comparative analysis of decision trees vis-a-vis other computational data mining techniques in automotive insurance fraud detection. J. Data Sci., 10: 537-561.
15. Paefgen, J., T. Staake and F. Thiesse, 2013. Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. Decis. Support Syst., 56: 192-201.
16. Guelman, L. and M. Guillén, 2014. A causal inference approach to measure price elasticity in automobile insurance. Expert Syst. Applic., 41: 387-396.