International Journal of Computer & Information Technologies (IJOCIT)
www.ijocit.ir & www.ijocit.org
ISSN = 2345-3877

Survey of Detecting Fraud in Automobile Insurance Using Data Mining Techniques

Leila Goleiji *[1] ,Mohammad Jafar Tarokh[2]

[1] *Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*
[2] *Associate professor, IT Group - Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran*

goleijileila@gmail.com

Abstract:- Insurance companies are exposed various types of fraud And each year suffer substantial losses. These companies are looking for ways to review the claims of its customers to detect and prevent fraud and financial abuse from their insurance. Nowadays, data mining methods are used to discover and extract knowledge from dataset . The utilization of this methods can be help to fraud detection in insurance industry. The primary goal of this survey is fraud detection methods of automobile insurance using data mining approach during the past 19 years. Findings in this study indicate that data mining methods like logistic models, Bayesian and decision trees have been applied most widely to provide preliminary solutions to the difficulties intrinsic in the identification and categorize of fraudulent automobile insurance data.

Keywords: Insurance fraud, Data mining, Fraud detection , Automobile insurance.

1. Introduction

Insurance companies are among the influential firms in the economic environment of a country [1]. Fraud is one of the biggest problems of insurance sector, which causes significant amounts of financial loss. In insurance literature, fraud refers to a deliberate framing of an unreal claim, loss declaration beyond reality or any other plot to obtain money more than what the insurer deserves [2]. Fraud has adverse effects on the company in various ways including financial aspects, reputation, organizational fame, and psychological;-social functions [3].

Despite great improvement in the methods of detecting such plots, the cost derived from such frauds to the companies is still increasing. Fraud

in insurance may occur in various stages and by various entities such as new insurers, existing insurers, affected third parties, or even experts who provide services to insurers [4]. Fraud in insurance may take various forms and may take place around us much more times than what we expect every day [5].

In this way, untrue claims and the possibility of policyholders' fraud seeking money from insurance companies lead to the fear of risk and uncertainty among the companies; in turn, they may increase their premiums and, consequently, everyone will be the victim of a minority behavior [6].

In this paper, we try to explore the literature on the fraud detection in automobile insurance from 1997 to 2016. The techniques used to fraud detection in automobile industry in each paper are described separately, and finally the more common ones during the past 19 years are identified and Be introduced.

2. Insurance Fraud Detection and Data Mining Classification Framework

In this portion, classification framework is propound for the current manuscript regarding applications of data mining techniques to insurance fraud detection (IFD). classification framework is on the basis of literature review papers published previously [7] as well as present knowledge regarding nature of data mining research[8].

Our suggestion is insurance fraud detection in automobile insurance. The usual data mining techniques in the insurance fraud detection are classified into six data mining application classes, including classification, clustering, prediction, outlier detection, regression, and visualization. We provide a brief description of the six data mining application classes.

Classification: Classification establishes and uses a model in order to predict the categorical labels of unknown objects aiming at differentiation between objects of various classes. These categorical labels are predefined, discrete and unorganized[9],[10].

Clustering: Clustering is applied to divide objects into conceptually meaningful groups (clusters), while the objects in a group being similar to one another but very dissimilar to the objects in other groups. Clustering is, also, known as data segmentation and regarded as a type of unsupervised classification[9],[10].

Prediction: based on the patterns of a data set, Prediction estimates numeric and ordered future values [11],[12]. Han and Kamber note that, for the purpose of prediction , the attribute for which the values being predicted is continuous-valued rather than categorized (discrete-valued and unordered)[9].

Outlier detection: Outlier detection is applied to measure the "Distance" among data objects for the purpose of detecting those objects which are different from, or inconsistent with, the remaining data set[9]: Data appeared to have various characteristics compared to the rest of the population are known as outliers[13]. Yamanishi et al. mention that the problem of outlier detection is one of the most fundamental issues in data mining[14].

Regression: Regression is a statistical methodology applied to manifest the relationship between one or more independent variables and one dependent variable [9]. A lot of experimental studies have used logistic regression as a benchmark. The regression technique is to be undertaken typically by use of

© 2016, IJOCIT All Rights Reserved

such mathematical methods as logistic regression and linear regression, and used to the detect various types of fraud detection[15-19].

Visualization: Visualization refers to the easily understand presentation of data and methodology which turns complicated data characteristics clear patterns or relationships uncovered in the data mining process[20],[21]. Eick and Fyock report that researchers at Bell and AT&T Laboratories have utilized the pattern detection capabilities of the human visual system through establishing a suite of tools and applications which encode data flexibly using color, position, size and other visual characteristics. Through clear presentation of data or functions, Visualization is best used [21].

3. Fraud Detection in Automobile Insurance via Data mining Techniques

In this paper, we try to explore the literature on the fraud detection in automobile insurance from 1997 until 2016. Table 1 provides us with the data mining program classes and data mining techniques utilized in 37 studies on the fraud detection in automobile insurance.

Table 1. Fraud detection in automobile insurance by applied Techniques.

Datamining application class	Data mining technique	Reference
Classification	Neural network, naïve Bayesian, Decision trees	[22]
	Principal component analysis of RIDIT(PRIDIT)	[23]
	Fuzzy logic	[24]
	Bayesian belief network ,Logistic model	[25]
	Self-organizing map	[26]
	Naïve Bayes	[27]
	Genetic Algorithm	[28]
	Fuzzy Expert System	[29]
	Logistic model, Bayesian belief network ,Neural networks, K-nearest neighbor, Naïve Bayes, svm, Decision trees.	[30]
	support vector machine, Genetic programming	[31]
	SVM , Naive Bayes tree, SVM-RFE (recursive feature elimination), Decision tree	[32]
	Decision trees ,Consolidated Trees	[33]
	Decision Tree ,Naïve Bayesian	[34]
	Ensemble neural network, Neural network	[35]
	Fuzzy logic	[36]
	Decision tree	[37]
	Fuzzy DEMATEL, Intuitionist fuzzy number, ELECTRE-TRI	[38]
	Gradient Boosting	[39]

© 2016, IJOCIT All Rights Reserved

	Decision tree(c4.5,id3,chaid)	[40]
	Decision tree ,Naïve Bayesian	[41]
	Naïve Bayesian, Decision tree, support vector machine	[42]
	Neural network	[43]
	Logistic model, Naïve Bayesian, Decision tree	[44]
	Decision tree, Svm, ANN	[58]
Prediction	Evolutionary algorithms, Cultural algorithms	[45]
	Social network analysis, Iterative Assessment Algorithm (IAA)	[46]
	GA-Kmeans, MPSO-Kmeans	[47]
Regression	Probit	[48]
	Logistic	[49]
	Probit	[50]
	Logistic	[51]
	Logistic	[52]
	Logistic	[53]
	Logistic	[54]
	Logistic	[55]
	Logistic	[56]
	Logistic	[43]
	Logistic	[57]

According to Table 1, the number of the articles that used similar techniques in order to detect fraud in automobile insurance (Table 2) is been extracted. According to Table 2, the techniques Logistic model (11 articles of the 37 articles) , Decision Tree (10 articles of the 37 articles) and Naïve Bayes (6 articles of the 37 articles) and Neural network (6 articles of the 37 articles) have been used in greater number of studies.

Table 2. Distributing Articles Using Data Mining Techniques.

No.	Technique	Number of papers
1	Logistic model	11
2	Decision Tree	10
3	Naïve Bayes	6
4	Svm	4
5	Probit model	2
6	Bayesian belief	2
7	Self-organizing map	1
8	Social network analysis	1
9	Neural network	6
10	Consolidated Trees	1
11	Cultural algorithms	1
12	ELECTRE-TRI	1
13	Evolutionary algorithms	1
14	Fuzzy DEMATEL	1

© 2016, IJOCIT All Rights Reserved

15	Fuzzy logic	2
16	Genetic programming	1
17	Gradient Boosting	1
18	Intuitionistic fuzzy number	1
19	Iterative Assessment Algorithm (IAA)	1
20	K-nearest neighbor	1
21	Principal component analysis of RIDIT	1
22	ensemble neural network	1
23	SVM (recursive feature elimination)	1
24	Genetic Algorithm	1
25	Fuzzy Expert System	1
26	GA-Kmeans	1
27	MPSO-Kmeans	1

Table 3 shows the usage percentage of each data mining application class in the reviewed papers. The highest rate was for Classification with 63% of usage in the reviewed studies. Regression and Prediction were of the lowest priorities in these studies with 29% and 8% of usage respectively.

Table 3. Distribution Articles using data mining application class.

Data mining application class	Number of papers	Percentage
Classification	24	63%
Prediction	3	8%
Regression	11	29%

Table 4 shows the number of papers in the field of automobile insurance fraud from 1997 until 2016. The number of papers in recent years indicates that the fraud detection in automobile insurance is still a challenging job for insurance company so that about 36% of the studies have been performed during the past five years.

Table 4. Distribution of Articles by Propagation Year

Year	Number of papers
1997	2
1998	1
1999	2
2000	1
2001	1
2002	5
2003	1
2004	1
2005	5
2007	2
2008	1
2010	1

© 2016, IJOCIT All Rights Reserved

2011	4
2012	4
2015	5
2016	1

4. Conclusion

In this study, the techniques for fraud detection in automobile insurance proposed by scientific studies from 1997 to 2016 (almost all papers in this regard) were reviewed. According to the analyses, we found out that Classification of data mining application class was used by 63% of the studies and is regarded as the best class of detection in automobile insurance. Among Classification techniques, Logistic Model, Decision Tree, and Naïve Bayes had the highest rate of usage.

References

[1] Azam Mohammad Beigi A, "An Introduction to Insurance Fraud: A Case of Third Party Insurance", Current Matters in Insurance World, 2005, pp. 25-36.

[2] Gill K M, Woolley K A, Gill M, "insurance fraud : the business as a victim", crim at work 1, 1994 ,pp.73-82.

[3] Rahimian N , Akhond Zadeh M, "Role of Internal Auditors in Preventing and Fraud Detection", Official Accountant,2011, pp. 20-44.

[4] Terisa R, " Improving the Defense Lines: The Future of Fraud Detection in the Insurance", SAS Global Forum,2010.

[5] Wilson J, "An Analytical Approach To Detecting Insurance Fraud Using Logistic Regression", Journal of Finance and Accountancy, 2003, pp.1-15.

[6] Varharami V, "Effective factors on fraudulent claims in automobile insurance in Iran", Insurance research paper, 2010, pp.145-160.

[7] Ngai E, Hu Y, Wong Y, Chen Y & Sun X, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", Decision Support Systems, 2011, pp.559-569.

[8] Ahmed S R,(2004). "Applications of data mining in retail business", Information Technology: Coding and Computing, pp. 455-459.

[9] Han J , Kamber M, "Data Mining: Concepts and Techniques", Second ed, Morgann Kaufmann Publishers,2006.

[10] Tan P, Steinbach M , Kumar V, "Introduction to Data Mining", First ed.Addison-Wesley Longman Publishing Co., Inc, 2005.

[11] Berry M J , Linoff G S, "Data Mining Techniques: for Marketing", Sales and Customer Relationship Management, Second ed.Wiley, New York, 2004.

[12] Agyemang M , Barker K , Alhajj R, "A comprehensive survey of numeric and

© 2016, IJOCIT All Rights Reserved

symbolic outlier mining techniques", Intelligent Data Analysis, 10,2006, pp.521–538.

[13] Yamanishi K, Takeuchi J, Williams G & Milne P, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms", Data Mining and Knowledge Discovery, 8: 2004, pp.275–300.

[14] Agresti A, "Categorical Data Analysis", Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1990.

[15] Duda R O, Hart P E ,Stock E G, "Pattern Classification", Wiley, New York,1996.

[16] Sharma S, "Applied Multivariate Techniques", Wiley, New York,1996.

[17] Viaene S, Derrig R A, Baesens B , Dedene G, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection", The Journal of Risk and Insurance ,2002, pp.373–421.

[18] Webb A,"Statistical Pattern Recognition", Arnold, London,1999.

[19] Shaw M J, Subramaniam C, Tan G W ,Welge M E, "Knowledge management and data mining for marketing", Decision Support System ,2001, pp.127–137.

[20] Turban E, Aronson J E, Liang T P , Sharda R,"Decision Support and Business Intelligence Systems", Eighth ed, Pearson Education,2007.

[21] Eick S G , Fyock D E, Visualizing corporate data, AT&T Technical Journal 75, 1996, pp.74–86.

[22] Viaene S, Dedene G , Derrig R A, "Auto claim fraud detection using Bayesian learning neural networks", Expert Systems with Applications 29,2005, pp.653–666.

[23] Brockett P L, Derrig R A , Golden L L, "Fraud classification using principal component analysis of RIDITS", The Journal of Risk and Insurance 69,2002, pp. 341–371.

[24] Pathak J, Vidyarthi N , Summers S L, "A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims", Managerial Auditing Journal 20,2005, pp.632–644.

[25] Bermúdez L, Pérez J M, Ayuso M, Gómez E, Vázquez F J & Bayesian Dichotomous A,"Model with asymmetric link for fraud in insurance", Insurance: Mathematics and Economics 42,2008, pp.779–786.

[26] Brockett P L, Xia X , Derrig R A, "Using Kononen's self-organizing feature map to uncover automobile bodily injury claims fraud", The Journal of Risk and Insurance 65,1998, pp.245–274.

[27] Viaene S, Derrig R A , Dedene G, "A case study of applying boosting naive Bayes to claim fraud diagnosis", IEEE Transactions on Knowledge and Data Engineering 16,2004, pp.612–620.

[28] lee B , Kim M, "Application of genetic algorithm to automobile insurance for selection of classification variables:the case of Korea", Annual Meeting of the American Risk and Insurance Association,1999.

[29] Taghavi Fard S M , Jafari Z, "Detecting Deception in Vehicle Insurance Using Fussy Expert System", Journal of Information Technology,2015, pp. 239-258.

© 2016, IJOCIT All Rights Reserved

[30] Viaene S, Derrig R A, Baesens B , Dedene G, "A Comparison of State - of - the - Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection", Journal of Risk and Insurance, 69,2002, pp.373-421.

[31] Vasu M , Ravi V, "A hybrid under-sampling approach for mining unbalanced datasets: applications to banking and insurance", International Journal of Data Mining, Modelling and anagement, 3,2011, pp.75-105.

[32] Farquad M, Ravi V , Raju S B, "Analytical CRM in banking and finance using SVM: a modified active learning–based rule extraction approach", International Journal of Electronic Customer Relationship Management, 6,2012, pp. 48-73.

[33] Pérez J M, Muguerza J, Arbelaitz O, Gurrutxaga I , Martín J I, "Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance", in Pattern Recognition and Data Mining, ed: Springer,2005, pp. 381-389.

[34] Bhowmik R, "Detecting auto insurance fraud by data mining techniques", Journal of Emerging Trends in Computing and Information Sciences, 2,2011, pp. 156-162.

[35] Xu W, Wang S, Zhang D , Yang B, (2011) "Random Rough Subspace Based Neural Network Ensemble for Insurance Fraud Detection", in Computational Sciences and Optimization (CSO), Fourth International Joint Conference on, pp. 1276-1280

[36] Bordoni S & Facchinetti G, (2001) "Insurance fraud evaluation a fuzzy expert system", IEEE International Fuzzy Systems Conference.

[37] D'Arcy S P, "Predictive modeling in automobile insurance: a prelim-inary analysis", in World Risk and Insurance Economics Congress, SaltLake City, Utah,2005.

[38] Taghiloo E, Azar A, Nasiri E , Ghaedrahmati H, "A New Model for Insurance Fraud Detection in Car Accidents Using a Combined Fuzzy Dematel and ELECTRE-TRI Approach", International Journal of Business and Social Science, 3,2012, pp.122-136.

[39] Guelman L, "Gradient boosting trees for auto insurance loss cost modeling and prediction", Expert Systems with Applications, 39,2012, pp.3659-3667.

[40] Goleiji L,(2015), "Fraud Detection in Insurance Industry Using Decision Tree Algorithms: A Case Study of Automobile Insurance", Second Conference of Engineering Sciences Development. Tonekabon: Ayandegan Institute of Higher Education.

[41] Goleiji L , Tarokh M J,(2015) "Identification of Effective variables and Fraud Detection in Insurance industry using data mining techniques: A case study, Automobile insurance", The 7th conference of Electrical engineering, Gonabad, Iran.

[42] Goleiji L , Tarokh M, (2015) "Fraud Detection in Insurance using data mining algorithms of decision tree, naïve Bayes, and support vector machine: A case study: automobile insurance", The 7th

© 2016, IJOCIT All Rights Reserved

conference of electrical engineering, Gonabad, Iran.

[43] Caldeira A M, Gassenferth W, Machado M A S , Santos D J ,(2015) "Auditing Vehicles Claims using Neural Networks", Information Technology and Quantitative Management (ITQM), Procedia Computer Science, 55, pp.62 – 71.

[44] Firoozi M, Shakoori M, Kazemi L ,Zahedi S, "Fraud detection in automobile insurance using data mining methods", Insurance research journal,2011, pp. 103-128.

[45] Sternberg M , Reynolds R G, "Using cultural algorithms to support re-engineering of rule-based expert systems in dynamic performance environments: a case study in fraud detection", Evolutionary Computation, IEEE Transactions on, 1,1997, pp.225-243.

[46] Šubelj L, Furlan S , Bajec M, "An expert system for detecting automobile insurance fraud using social network analysis", Expert Systems with Applications, 38,2011, pp.1039-1052.

[47] Liu J , Chen C,(2012) "Application of Evolutionary Data Mining Algorithms to Insurance Fraud Prediction", 4th. International Conference on Machine Learning and Computing, pp. 22-17.

[48] Pinquet J, Ayuso M, Guillén M, "Selection bias and auditing policies for insurance claims", The Journal of Risk and Insurance 74,2007, pp.425–440.

[49] Crocker K J , Tennyson S, "Insurance fraud and optimal claims settlement strategies", Journal of Law and Economics, 45,2002, pp.469–507.

[50] Belhadji E B, Dionne G , Tarkhani F, "A model for the detection of insurance fraud", Geneva Papers on Risk and Insurance, Issues and Practice,2000 , pp.517-538.

[51] Weisberg H I , Derrig R A, "Quantitative methods for detecting fraudulent automobile bodily injury claims", Risques, 35,1998, pp.75-101.

[52] Wilson J, "An Analytical Approach To Detecting Insurance Fraud Using Logistic Regression", Journal of Finance and Accountancy,2003 , pp.1-15.

[53] Caudill S B, Ayuso M, Guillén M, "Fraud detection using a multinominal logit model with missing information", The Journal of Risk and Insurance, 72,2005, pp.539–550.

[54] Tennyson S, Salsas Forn P, "Claims auditing in automobile insurance: fraud detection and deterrence objectives", The Journal of Risk and Insurance, 69,2002, pp.289–308.

[55] Viaene S, Ayuso M, Guillen M, Van Gheel D & Dedene G, "Strategies for detecting fraudulent claims in the automobile insurance industry", European Journal of Operational Research, 176,2007, pp.565-583.

[56] Artís M, Ayuso M & Guillén M, "Modelling different types of automobile insurance fraud behaviour in the Spanish market, insurance", Mathematics and Economics, 24,1999, pp.67–81.

[57] Artís M, Ayuso M & Guillén M, "Detection of automobile insurance fraud with discrete choice models and misclassified claims", The Journal of Risk and Insurance, 69,2002, pp.325–340

© 2016, IJOCIT All Rights Reserved

[58] Amira Kamil Ibrahim H, Ajith A, (2016)," Modeling Insurance Fraud Detection Using Imbalanced Data Classification", Advances in Nature and Biologically Inspired Computing,Advances in Intelligent Systems and Computing, , ,pp.117–127.

© 2016, IJOCIT All Rights Reserved