WWJMRD 2017; 3(8): 178-181
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
UGC Approved Journal
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

Ananthi Sheshasaayee
Research Supervisor
PG & Research Department of Computer Science, Quaid-E-Millath Government College for Women Anna Salai, Chennai, Tamilnadu, India

Surya Susan Thomas
Research Supervisor
PG & Research Department of Computer Science, Quaid-E-Millath Government College for Women Anna Salai, Chennai, Tamilnadu, India

Correspondence:
Surya Susan Thomas
Research Supervisor
PG & Research Department of Computer Science, Quaid-E-Millath Government College for Women Anna Salai, Chennai, Tamilnadu, India

A Study of Hybrid Learning Methodologies in Insurance Fraud Detection Techniques

Ananthi Sheshasaayee, Surya Susan Thomas

Abstract

Insurance fraud is a false misinterpretation of a matter of fact to obtain a financial gain from an insurance claim process. This result when a claimant attempts to achieve some financial boost which he/she is not supposed to obtain. Nowadays, the scenario has changed in such a way that in a developing country like India, the realization to get insured has substantially increased over the period of twenty to thirty years. The increase in insurance claims also led to an increase in the fraudulent claims which led to the necessity to detect these claims at the earliest so as to reduce financial losses to the insurance companies. Data mining methods helps in detecting frauds to a great extent. An overview of the supervised, unsupervised and hybrid learning methodologies for insurance fraud detection is discussed in this paper.

Keywords: Insurance, Fraud detection, Datamining methods, Supervised, Unsupervised, Hybrid learning

Introduction

Insurance is a procedure by which an insurance company or state assures a compensation for specified loss, damage, illness or death in return for payment of a specified premium. In other words, insurance is a plan providing protection against a possible eventuality. Insurance in India is broadly classified into life, motor, health, travel, home, rural, commercial and business insurance.The Insurance Regulatory and Development Authority, an agency of the Government of India, is the regulatory body for the insurance sector's supervision and development in India. This agency was established in the year 1999 by the Indian Government, for two significant reasons, to protect the interest of the policy holders and for the promotion of the entire insurance sector right from the approach adopted by the existing insurance companies towards their shareholders to the eradication of the shortcomings of the industry. [1]

Intention to deceive, conceal or distort relevant information affecting benefits or derive undue benefits from Insurance industry is known as Insurance Frauds. Fraudulent claims costs crores of rupees annually. Insurance crimes range in severity, from slightly exaggerating claims to deliberately causing accidents. [2] Fraudulent activities affect the insurance providers as well as the policy holders equally since it is a financial loss either by the disposal of funds for fraudulent claims or by the increase in premium amount which has to be incurred by the customer. Insurance fraud is a significant problem, and both government and other organizations should make intense efforts to deter such activities. As automation takes place in all sectors of business, fraud detection also achieved a new paradigm with the introduction of data mining concept. Data mining offers a backing in detecting frauds efficiently and effectively.

Data Mining

Data Mining is the process of selecting, exploring and analyzing large amounts of data to mine previously unknown fact, which might prove to be a solution to the existing problem. In insurance industry, data mining and business intelligence work hand in hand to detect such fraudulent activities. Data Mining is the intrinsic part of the integrated information technology software that is usually referred as 'Business Intelligence'. These Information technology systems commences with the data warehousing, followed by Online analytical processing(OLAP) and concludes with the data mining methods. [3] This paper confines to the study and analysis of the various supervised, unsupervised and hybrid data mining methods employed to detect health insurance frauds. It is not possible to fully expunge frauds in this sector but can be reduced with the help of data mining methods and other business solutions. Figure:1 shows the data mining process for a business solution. [4]

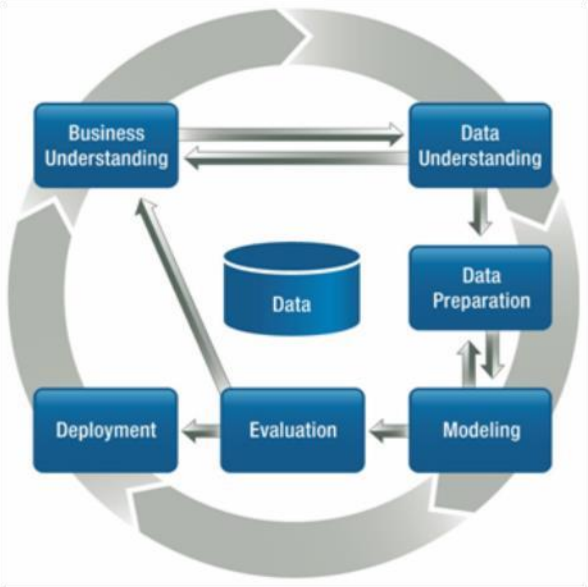

Fig. 1: Data Mining & Business Intelligence

Detecting Frauds

The traditional way of detecting fraud is the procedure which mainly comprises of the auditing of hundreds of claims manually [5]. But this takes a lot of time and energy, consequently led to the lag in the clearance of the claims which affects both the insurer and the policy holder. Automated business process and growth of computerized systems has led to sophisticated methods to detect fraud and abuse. Figure 2 illustrates the process in which a claim is investigated and fraud is detected, if any.

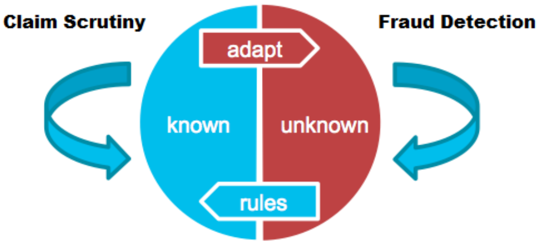

Fig. 2: Claim Investigation & Fraud Detection

Knowledge Discovery from Databases (KDD) emerged by combining information technology software systems and statiscal knowledge. Datamining is the spine of KDD. Data mining helps these insurance companies to extract useful information or knowledge from thousands of claims to form a smaller class of the doubtful claims for further scrutiny for fraud and abuse. [6]

Frauds are of different types in healthcare, mainly comprises of:

a) Billing services not rendered
b) Upcoding
c) Duplicate claims
d) Unbundling
e) Kickbacks

Supervised Learning Methods

Supervised learning methodologies works with the help of training data which includes both the input and the desired results. In some cases, correct results are identified and are given as input to the model during the learning process. Building up of a proper training validation and test set are very important. Supervised learning is usually fast and precise. [7] Decision tree, classification, regression are the major examples of supervised learning in insurance fraud detection domain.

Unsupervised Learning Methods

Unsupervised learning is a methodology where the training set is not available and finding patterns or structure in the data has to be done independently. This method typically analyses an insurance claim's attributes in relation to other claims and discover how they are related to or differ from each other. Algorithms are ported to the devices to explore and present the interesting pattern in the data. Some known examples of this learning are k-means for clustering techniques, outlier detection and Apriori algorithm for association rule learning problems. [7]

Hybrid Learning Methods

This methodology uses mainly unlabeled and a small amount of labelled input data. The usage of some labelled data can greatly enhance the efficiency of unsupervised learning tasks. [8] The model must learn the structure to organize the data as well as make predictions. The actual problem resides in between supervised and unsupervised learning. A study followed a three-step procedure for insurance fraud detection. It applied unsupervised clustering methods on insurance claims and modelled a variety of labelled clusters. Then they used an algorithm based on a supervised classification tree and discovered rules for the allocation of each record to clusters. They generated the most effective rules for future identification of fraud behaviors. [9]

Importance of Hybrid learning

Labeled data are difficult to obtain while unlabeled data are plenty, but inefficient in fraud detection therefore hybrid learning is a good solution to minimize human effort and improve accuracy. [10]

The aim of hybrid learning is to explore the combined effort of labelled and unlabeled data which might change the learning environment and formulate algorithms for effective fraud detection. This learning is interesting in machine learning and data mining concepts because it can make use of the unlabeled data to improvise supervised learning which are limited and expensive. Semi-supervised learning models such as self-training mixture models-training and Multiview learning, graph-based methods and semi-supervised support vector machines are the existing ones.

Methods in Hybrid learning

Because of the scarcity of the supervised or labelled data, hybrid learning models make strong model predictions. So, one should carefully choose a model which best fits the problem structure. Commonly deployed hybrid learning methods in this field are

- Generative Model
- Graph based Model
- Self-Training
- Employing Class Proportion Model

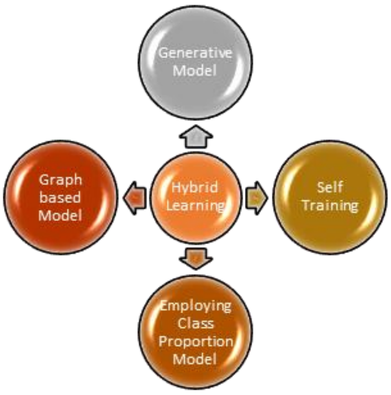

Fig. 3: Hybrid learning models

- **Generative Models**

It is one of the oldest hybrid learning methods. This model is used for randomly generating observable data values, typically data with hidden parameters.

- **Graph based models**

Graph-based hybrid methods is a graph where the nodes are labelled and unlabeled examples in the dataset, and the edges (may be weighted) reflect the similarity of examples. This method assumes label smoothness in the graph. Graph methods are nonparametric, discriminative and transudative in nature. [11]

- **Self-Training**

Self-training is a technique used in hybrid learning. In self-training, a classifier is first trained with the small amount of labelled data. It is then used to classify the unlabeled data. The most confident unlabeled points, with their predicted labels are then incorporated to the training set. The classifier is retrained and the procedure is repeated.

- **Employing Class Proportion Knowledge**

It has been studied that constraining the class proportions on unlabeled data can be important for semi-supervised learning. Various hybrid learning methods have been observed to use some form of class proportion constraints. The desired class proportions are either obtained as an input, which reflects domain knowledge, or estimated (by frequency or with smoothing) from the class proportions in the labelled dataset. [11]. by engaging these models of hybrid learning, a very beneficial model for efficient fraud detection can be drafted which could stimulate the financial gain of the insurance provider and proportionally reducing the burden on the customer.

Related Works

A number of researchers have developed hybrid methods by combining supervised and unsupervised methods. Unsupervised method proceeded by supervised method usually give way to discovery of knowledge in a hierarchical manner.

Rashidian et al applied supervised and unsupervised datamining techniques to detect healthcare frauds. SVM, neural networks, genetic algorithms, and decision trees were used from supervised learning and clustering, outlier detection and association rules were used from the unsupervised learning methodology. [12] They were able to retrieve better results by the hybrid learning method than by using any one methodology.

Anuradha et al integrates SVM (Support Vector Machine) and ECM (Evolving Clustering Method) in health insurance field for fraud detection. They used SVM algorithm was used for classification and EVM algorithm was used for clustering. [13] Applying these two methods, the system was trained to draw a boundary between legitimate and fraudulent claims with more accuracy.

Peng et al constructed a model to detect suspicious health care frauds from large databases using clustering techniques. [14] They applied two clustering methods namely SAS EM and CLUTO to health insurance dataset and compares their capability. CLUTO was faster but SAS EM gave more useful clusters.

Williams and Huang combined clustering algorithms and decision trees to detect insurance subscribers' fraud. [15] A three-step "divide and conquer" procedure was used to find the solution.

Abraham et al introduced two hybrid ideas for modelling Intrusion Detection System (IDS). Decision trees and SVMs are joined as a hierarchical hybrid intelligent model and an ensemble approach connecting the main classifiers. [16] This hybrid fraud detection model incorporates the classifiers and other hybrid machine learning approaches to bring out maximum fraud detection accuracy and reduce computational intricacy.

Zhang et al overviewed the advanced supervised machine learning and natural language processing techniques to the problem of detecting anomalies in financial reporting documents. [17]

Kose et al framed an innovative framework to expose fraudulent cases independent of the actors and entities involved in the claims. Interactive machine learning incorporating expert knowledge in an unsupervised setting was applied. They used pairwise comparison method of Analytical Hierarchical Processing (AHP) for weighing the actors and attributes and Expectation Maximization (EM) for clustering similar actors. Two stage data warehousing for proactive risk calculations, visualization tools for effective analyzing and Z-score and standardization to formulate the risks were also employed. [18] The framework which was named eFAD suite effectively handles the fragmented nature of the abnormal behaviors.

Major et al used hybrid learning in EFD (Electronic Fraud Detection). They used knowledge discovery techniques on two levels. [19] First, integrating knowledge with statistical information assessment and second machine learning was used to mould new guidelines and highlight the identification process.

Conclusion

Supervised learning methods applied to detect frauds gave accurate results but to access labeled data is expensive and limited. In order to overcome this scenario, unsupervised learning methods were introduced and this too had limitations of training the data which was time consuming and results might not be accurate as the supervised learning. With the induction of the hybrid learning method, which incorporates the supervised and unsupervised learning methodologies, fraud detection in insurance industry have become easier and cost-effective. It gives an insight to improvise the existing fraud detection methodologies, not only in insurance sector but also in other business sectors.

References

1. IRDAI;" History of Insurance in India", Ref: IRDA/GEN/06/2007, July 2007.
2. Ngufor, Che, and A. Wojtusiak. "Unsupervised labeling of data for supervised learning and its application to medical claims prediction." *Computer Science*14 (2013).
3. Joudaki, Hossein, et al. "Using data mining to detect health care fraud and abuse: a review of literature." *Global journal of health science* 7.1 (2014): 194-202
4. http://www.rosebt.com/blog/category/enterprise %20data%20warehousing%20platforms/2
5. Ortega, Pedro A., Cristián J. Figueroa, and Gonzalo A. Ruz. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile." *DMIN* 6 (2006): 26-29.
6. Faseela, V. S., and P. Thangam. "A Review on Health Insurance Claim Fraud Detection." *International Journal of Engineering Research Science (IJOER)* 1 (2015).
7. http://www.astro.caltech.edu/~george/aybi199/Donalek_Classif.pdf
8. https://www.nesta.org.uk/sites/default/files/machines_that_learn_in_the_wild.pdf
9. Major, John A., and Dan R. Riedinger. "EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud." *Journal of Risk and Insurance*69.3 (2002): 309-324.
10. Punam Devidas Bagul, Sachin Bojewar, Ankit Sanghavi." Survey on Hybrid Approach for Fraud Detection in Health Insurance". International Journal of Innovative Research in Computer and Communication EngineeringVol 4,Issue 4, April(2016)
11. Zhu, Xiaojin. "Semi-supervised learning literature survey." (2005).
12. Joudaki, Hossein, et al. "Using data mining to detect health care fraud and abuse: a review of literature." *Global journal of health science* 7.1 (2014): 194-202.
13. Rawte, Vipula, and G. Anuradha. "Fraud detection in health insurance using data mining techniques." *Communication, Information & Computing Technology (ICCICT), 2015 International Conference on*. IEEE, 2015
14. Peng, Yi, et al. "Application of clustering methods to health insurance fraud detection." *2006 International Conference on Service Systems and Service Management*. Vol. 1. IEEE, 2006.
15. Williams, Graham J., and Zhexue Huang. "Mining the knowledge mine." *Australian Joint Conference on Artificial Intelligence*. Springer Berlin Heidelberg, 1997.
16. Peddabachigari, Sandhya, et al. "Modeling intrusion detection system using hybrid intelligent systems." *Journal of network and computer applications*30.1 (2007): 114-132.
17. Seemakurthi, Prasad, Shuhao Zhang, and Yibing Qi. "Detection of fraudulent financial reports with machine learning techniques." *Systems and Information Engineering Design Symposium (SIEDS), 2015*. IEEE, 2015.
18. Kose, Ilker, Mehmet Gokturk, and Kemal Kilic. "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance." *Applied Soft Computing* 36 (2015): 283299
19. Major, John A., and Dan R. Riedinger. "EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud." *Journal of Risk and Insurance*69.3 (2002): 309-324.