ISSN:0975-9646

Analysis of Fraud Detection Mechanism in Health Insurance Using Statistical Data Mining Techniques

Pravin R. Bagde[1], Manoj S. Chaudhari[2]

[1]PG Student, Department of CSE, PBCE Nagpur, India

[2]Professor & HOD, Department of CSE, PBCE Nagpur, India

***Abstract* - Health insurance fraud increases the disorganization and unfairness in our society. Health care fraud leads to substantial losses of money and very costly to health care insurance system. It is horrible because the percentage of health insurance fraud keeps increasing every year in many countries. To address this widespread problem, effective techniques are in need to detect fraudulent claims in health insurance sector. The application of data mining is specifically relevant and it has been successfully applied in medical needs for its reliable precision accuracy and rapid beneficial results. This paper aims to provide a comprehensive survey of the statistical data mining methods applied to detect fraud in health insurance sector.**

***Keywords* - health insurance, fraud detection, data mining, statistical methods.**

I. INTRODUCTION

Insurance fraud is a significant and costly problem for both policyholders and insurance companies in all sectors of the insurance industry. India is one of the fastest growing economies in the world, has a burgeoning middle class, and has witnessed a significant rise in the demand for health insurance products. Over the last 10 years, the health insurance industry has grown at a capital annual compounded growth rate of around 20%. However, with the exponential growth in the industry, there has also been an increased incidence of frauds in the country.

Health Insurance fraud encompasses a wide range of illicit practices and illegal acts involving intentional deception or misrepresentation. Data mining has a tremendous impact in improving healthcare fraud detection system. Data mining has been applied to fraud detection in both the way i.e. supervised and non-supervised manner. Data mining techniques and its application for fraud detection in health care sector is described below.

II. DATA MINING TECHNIQUES & ITS APPLICATION

In today’s world there is huge amount of data stored in real world databases and this amount continues to grow fast. Traditional methods of detecting health care fraud and abuse are time-consuming and inefficient. So, there is a need for semi-automatic methods that discover the hidden knowledge in such database. Data mining is a core of the KDD process. Data mining automatically filtering through immense amounts of data to find known or unknown patterns bring out valuable new perceptions and make predictions. Data mining techniques has been used intensively and extensively by many healthcare organizations for fraud detection.

A. Data Mining Techniques

Data mining techniques tend to learn models from data. There are two approaches on learning the data mining models. One is supervised learning & second is unsupervised learning they are described below:

Supervised methods are usually used for classification and prediction objectives including traditional statistical methods such as regression analysis, discriminant analysis, neural networks, Bayesian networks and Support Vector Machine (SVM). Supervised learning methods attempt to discover the relationship between input variable and an output variable.

Unsupervised methods are usually used for description including association rules extraction such as Apriori algorithm and segmentation methods such as clustering and anomaly detection. Unsupervised learning methods are applied when no prior information of the dependant variable is available for use.

B. Application of Data Mining:

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions [2].

Data mining application can greatly benefit all parties involved in the health care industry. For example data mining can help healthcare insurers to detect fraud and abuse; health care organization can make customer relationship management decision; Physicians can identify effective treatments and best practices; and patients receive better and more affordable healthcare services.

III. Statistical Data Mining Techniques For Health Care Fraud Detection

The statistical data mining methods effectively consider big data for identifying structures (variables) with the appropriate predictive power in order to yield reliable and robust large-scale statistical models and analyses.

The net effect of excessive fraudulent claims is excessive billing amounts, higher per-patient costs, excessive per-doctor patients, higher per-patient tests, and so on. This excess can be identified using special analytical tools.

Provider statistics include; total number of patients, total amount billed, total number of patient visits, per-patient average visit numbers, per-patient average billing amounts, per-patient average medical test costs, per-patient average medical tests, per-patient average prescription ratios (of specially monitored drugs) and many more.

Existing research approaches for health care fraud detection can be divided into three classes: supervised methods, unsupervised methods, and hybrid methods.

A. Supervised Fraud Detection Methods

There are several supervised fraud detection methods such as: Bayesian Networks, Neural Networks (NNs), Decision Trees, and Fuzzy Logic. NNs and decision trees are the most popular fraud detection methods because of their high tolerance of noisy data and huge data set handling. [3]

Bayesian Networks provide a graphic model of causal relationships on which class membership probabilities (Han et al. 2000) are predicted, so that a given instance is legal or fraud (Prodromidis, 1999). Naïve Bayesian classification assumes that the attributes of an instance are independent; given the target attribute.

Multilayer perceptron (MLP) neural network is a widely used supervised technique in health care fraud detection because it has many advantages such as it can handle complex data structure especially non-linear relationship and it has high tolerance to noisy data [5].

Decision trees (DTs**)** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

There are some of the advantages of decision trees and these are: It is simple to understand and to interpret, able to handle both numerical and categorical data, able to handle multi-output problems.

If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret and possible way to validate a model using statistical tests.

That makes it possible to account for the reliability of the model. Performs well even if its assumptions are some what violated by the true model from which the data were generated.

B. Unsupervised Fraud Detection Methods

It finds “natural” grouping of instances given un-labeled data. Compared to supervised health care fraud detection methods, which centralized on MLP neural networks and decision trees, unsupervised health care fraud detection methods various a lot, ranging from self-organizing map, association rules, clustering, to rule-based unsupervised methods.

Some experts has developed an expert system, called electronic fraud detection, to detect service providers’ fraud. This system based on unsupervised rule-based algorithm to scan health insurance claims in search of likely fraud.

EFD has applied rule-based methods on two levels. On the first level, EFD integrates expert knowledge with statistical information assessment to induce rules to identify cases of unusual provider behaviour. On the second level, these rules were validated by the set of known fraud cases. According to the validation results, fuzzy logic is used to develop new rules and improve the identification process.

When operating this system, each provider’s behaviour was measured first. Then it was compared to that of its peers (providers with the same organizational structure and specialty, and practicing in the same geographical area); if the provider stands out from the mainstream it was recognized as the suspicious fraudulent service provider.

In [7], a self-organizing map (SOM), a type of unsupervised neural network, was applied to general practitioners database to create an unbiased subdivision of general practitioners’ practices for the purpose of more effective monitoring of test ordering. The results indicated that after applying SOM, general practitioners were successfully classified into groups of various sizes, which reflect the nature and style of their practices.

The author also used association rules on insurance claims database to identify commonly related test. Therefore, if one or more of the tests in a claim has little association with the other tests, this claim would be identified as suspicious fraudulent claim. Association rules have given a new perspective to health care fraud detection problem, and the output rules can be utilized to facilitate other fraud detection methods.

In [11], SmartSifter, a system based on finite mixture model, is designed for on-line unsupervised outlier detection. SmartSifter uses a probabilistic model to represent the underlying data-generating mechanism. In the probabilistic model, a histogram is used to represent the probability distribution of categorical variables; for each bin of the histogram, a finite mixture model is used to represent the probability distribution of continuous variables.

When a new case is coming, SmartSifter updates the probabilistic model by employing an SDLE (Sequentially Discounting Laplace Estimation) and an SDEM (Sequentially Discounting Expectation and Maximizing) algorithm to learn the probability distributions of the categorical and continuous variables, respectively. A score is given to this new case, measuring how much the probabilistic model has changed since the last update. A high score indicates that this new case may be an outlier.

C. Hybrid methods

Hybrid methods, combining supervised and unsupervised methods, have been developed by a number of researchers. When an unsupervised method is followed by a supervised method, the objective is usually to discover knowledge in a hierarchical way. Williams and Huang [5] integrated clustering algorithms and decision trees to detect insurance subscribers' fraud. This procedure has been followed in some other unsupervised-supervised methods, but different algorithms were used in each step. For example, Williams [5] recommended the use of a genetic algorithm to generate rules such that extra freedom can be achieved, e.g., specification/revision of the rules by domain experts and the evolving of rules by statistical learning.

Hybrid methods of combining supervised and unsupervised methods also have been applied by some studies. Major and Riedinger [6] tested an electronic fraud detection program that compared individual provider characteristics to their peers in identifying unusual provider behavior. Unsupervised learning is used to develop new rules and improve the identification process. One study conducted a three step methodology for insurance fraud detection. They applied unsupervised clustering methods on insurance claims and developed a variety of (labeled) clusters. Then they used an algorithm based on a supervised classification tree and generated rules for the allocation of each record to clusters. They identified the most effective 'rules' for future identification of abusive behaviour.

D. Examples of statistical data analysis techniques are:

1. Data preprocessing techniques for detection, validation, error correction, and filling up of missing or incorrect data.
2. Calculation of various statistical parameters such as averages, quintiles, performance metrics, probability distributions, and so on. For example, the averages may include average length of call, average number of calls per month and average delays in bill payment.
3. Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
4. Computing user profiles.
5. Time-series analysis of time-dependent data.
6. Clustering and classification to find patterns and associations among groups of data.
7. Matching algorithms to detect anomalies in the behavior of transactions or users as compared to previously known models and profiles. Techniques are also needed to eliminate false alarms, estimate risks, and predict future of current transactions or users.

IV. Conclusion

The overall objectives of this survey paper is to analyse fraud detection mechanism using statistical data mining techniques because a fraud becomes more sophisticated and the volume of data grows, and that's why fraud becomes more difficult to recognise from bulk of data. The applications of statistical data mining methods to health care fraud detection are rewarding but highly challenging and therefore the main aim behind statistical approach is to improve fraud detection accuracy.

References

[1] Kirlidog, M., & Asuk, C. (2012). "A Fraud Detection Approach with Data Mining in Health Insurance". Procedia-Social and Behavioral Sciences, 62, 989-994. http://dx.doi.org/10.1016/j.sbspro.2012.09.168.

[2] Melih Kirlidoga, Cuneyt Asuk(2012) "A fraud detection approach with data mining in health insurance". Procedia - Social and ehavioral Sciences 62 (2012) 989 – 994.

[3] Jing Li & Kuei-Ying Huang & Jionghua Jin & Jianjun Shi"A survey on statistical methods for health care fraud detection" Health Care Manage Sci DOI 10.1007/s10729-007-9045-4

[4] RekhaBhowmik (2008) "Data Mining Techniques in Fraud Detection" Journal of Digital Forensics, Security and Law, Vol. 3(2).

[5] Williams, G. J., & Huang, Z. (1997). "Mining the knowledge mine. In Advanced Topics in Artificial Intelligence" (pp. 340-348). Springer Berlin Heidelberg. http://dx.doi.org/10.1.1.35.8596.

[6] Major, J. A., & Riedinger, D. R. (1992). EFD: A hybrid knowledge /statistical based system for the detection of fraud. *International Journal of Intelligent Systems, 7*(7), 687-703. http://dx.doi.org/doi:10.1111/1539-6975.00025.

[7] Qi Liu MiklosVasarhelyi (2013) "Healthcare fraud detection: A surveyand a clustering model incorporating Geo-location information" 29th WorldContinuous Auditing and Reporting Symposium (29WCARS), November 21-22, 2013, Brisbane, Australia.

[8] Amira Kamil Ibrahim Hassan and Ajith Abraham "Computational Intelligence Models for Insurance Fraud Detection: A Review of a Decade of Research" Journal of Network and Innovative ComputingISSN 2160-2174, Volume 1 (2013) pp. 341-347

[9] LutfunNaharLata, Israt Amir Koushika and SyedaShabnamHasan (2015) "A Comprehensive Survey of Fraud Detection Techniques" Inter Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Volume 10 – No.2, December 2015.

[10] Prasanna Desikan, Kuo-Wei Hsu &Jaideep Srivastava (2011) "Data Mining for Health Care Management" SAIM International Conference on Data Mining April 28-30, 2011 Hilton Phoenix East/Mesa Arizona USA.

[11] Bolton, R. J., & Hand, D. J. (2002). "Statistical fraud detection: A review. Statistical Science", 235-249. http://dx.doi.org/10.1214/ss/1042727940.

[12] Jing Li, Kuei-Ying Huang, Jionghua Jin &Jianjun Shi (2008) "A survey on statistical methods for health care fraud detection" Health Care Manage Sci (2008) 11:275–287 DOI 10.1007/s10729-007- 9045-4 Springer Science + Business Media, LLC 2007

[13] Li, J., Huang, K. Y., Jin, J., & Shi, J. (2008). "A survey on statistical methods for health care fraud detection". Health Care Management Science, 11(3), 275-287. http://dx.doi.org/ 10.1007/s10729-007-9045-4.

[14] Dan Ventura. Class Lecture, Topic: "SVM Example." BYU University of Physics and Mathematical Sciences, Mar. 12, 2009.

[15] Fashoto Stephen G., Owolabi Olumide, Sadiku J., Gbadeyan Jacob A,"Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm",Australian Journal of Basic and Applied Sciences, 7(8): 140-144, 2013 ISSN 1991- 8178.