www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-5, Issue-5, June 2017
Research Paper

Int. J. Sc. Res. in Network Security and Communication

ISSN: 2321-3256

An Improved Approch For Fraud Detection In Health Insurance Using Data Mining Techniques

Namrata Ghuse[1*], Pranali Pawar[2], Amol Potgantwar[3]

[1*]Professor, Computer Engineering Department, SITRC, Pune University, Nashik, India
[2]Student, Computer Engineering Department, SITRC, Pune University, Nashik, India
[3]HOD, Computer Engineering Department, SITRC, Pune University, Nashik, India

[*]*Corresponding Author: pranalipawar491@gmail.com, Tel.: +91-8149581110*

Received 15th May 2017, Revised 24th May 2017, Accepted 14th Jun 2017, Online 30th Jun 2017

***Abstract*—** Nowadays there is huge amount of data stored in real world databases and this amount continues to grow fast. The major use of anomaly or outlier detection is fraud detection. Health care fraud leads to substantial losses of money each year in many countries. Effective fraud detection is important for reducing the cost of Health care system. Fraud and abuse on medical claims became a major concern for health insurance companies last decades. Fraud involves intentional deception or misrepresentation intended to result in an unauthorized benefit. It is shocking because the incidence of health insurance fraud keeps increasing every year. Data mining which is divided into two learning techniques viz., supervised and unsupervised is employed to detect fraudulent claims. Basically random forest algorithm and logistics regression algorithm techniques are used for fraud detection in health insurance. Data mining automatically filtering through immense amounts of data to find known/unknown patterns bring out valuable new perceptions and make predictions.

***Keywords*—** Data Mining; Random Forest Algorithm; Health Insurance Fraud; Supervised; Unsupervised; Clustering; Logistics Regression Algorithm.

I. INTRODUCTION

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 [1,2,3,4,5].

Insurance fraud is a significant ant 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 [6,7,8,9,10].

Health Insurance fraud encompasses a wide range of implicit 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 [11,12,13].

The health insurance fraud claims are broadly classified under the following headings [14,15,16,17]:

- *Billing for services not rendered:* Billing insurance company for things that never happened. Example: Forging the signature of those involved in giving bills.
- *Upcoding of services***:** Billing insurance company for services that are costlier than the actual procedure that was done. Example: 45-minute session being billed as 60-minute session
- *Upcoding of items:* Billing insurance company for medical equipment that is costlier than the actual equipment. Example: Billing for power assisted wheelchair while giving the patient only the manual wheelchair.

© 2017, IJSRNSC All Rights Reserved

- *Duplicate claims:* Not submitting exactly the same bill, but changing some small portion like the date in order to charge insurance company twice for the same service rendered. Example: An exact copy of the original claim is not filed for the second time, but rather some portion like date is changed to get the benefit twice the original.
- *Unnecessary services:* Filing claims which in no way apply to the condition of a patient. Example: Patient with no symptoms of diabetes filing claim for daily usage of insulin injections.

II. LITERATURE SURVEY

A. J. Yan et. al. investigates case studies where text messaging has been exploited to deliver safety information and early warnings to users based on the availability of their location information. But these are to complex LBS structure.

B. Phua et. al. highlights fraud committed in insurance industry as one of the most studied in terms of the number of data mining-based fraud detection publications, existing four sub-groups of insurance fraud detection: home, crop, automobile and medical insurances. In an on-line discounting learning algorithm to indicate whether a case has a high possibility of being a statistical outlier in data mining applications such as fraud detection is used for identifying meaningful rare cases in health insurance pathology data from Australia's Health Insurance Commission (HIC).

C. Becker et. al. identify the effects of fraud control expenditures and hospital and patient characteristics on up coding, treatment intensity and health outcomes in the Medicare and Medicaid programs. Cox applied a fraud detection system based on fuzzy logic for analyzing health care provider claims.

D. Ross Gayleret. al. defines the professional fraudster, formalises the main types and sub types of known fraud and presents the nature of data evidence collected within affected industries. Within the business context of mining the data to achieve higher cost savings, this research presents methods and techniques together with their problems.

E. Qi Liu et. al. develop sophisticated antifraud approaches incorporating data mining, machine learning or other methods. This introduce some preliminary knowledge of U.S. health care system and its fraudulent behaviours, analyses the characteristics of healthcare data and reviews and compares currently proposed fraud detection approaches using healthcare data in the literature as well as their corresponding data pre-process methods. Also a novel healthcare fraud detection method including geo-location information is proposed.

F. Dr. YuZhanget. al. explore show predictive fraud detection systems developed using ICD-9 claims data will initially react to the introduction of ICD-10. Expert have developed a basic fraud detection system incorporating both unsupervised and supervised learning methods in order to examine the potential fraudulence of both ICD-9 and ICD-10 claims in a predictive environment. Using this system, users are able to analyze the ability and performance of statistical methods trained using ICD-9 data to properly identify fraudulent ICD-10 claims. This research makes contributions to the domains of medical coding, healthcare informatics, and fraud detection.

G. Guido Cornelis van Capelleveen proposed methodology enabled successful identification of fraudulent activity in several cases; however linking these identified incidents with irrefutable de jure fraud proved to be a difficult process. From 17 top suspicions analysed, expert reported eventually 12 of those to officials, a precision rate of approximately 71%. In the two interviews conducted with Medicaid Fraud Experts, experiences were gained on requirements for the design of the analytics and an effective implementation of the method. Experts found that outlier based predictors are not likely to succeed as fraud classification technology, though it explored an important role as decision supportive technology for resource allocation of fraud audits.

III. STATISTICAL DATA MINING TECHNIQUES

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.

© 2017, IJSRNSC All Rights Reserved

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. Bayesian Networks provide a graphic model of causal relationships on which class membership probabilities are predicted, so that a given instance is legal or fraud. 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.

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

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 behavior. 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.

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 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 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.

IV. PROPOSED SYSTEM

First step of the system was to collect raw data from Medicare institutes. Below are the key elements of the data user would be focusing upon –

a) Claim ID – Unique reference number to identify claims

b) Claim Purchase Date – Date when the claim was made

c) Classification Type – Whether the claim was fraudulent or genuine

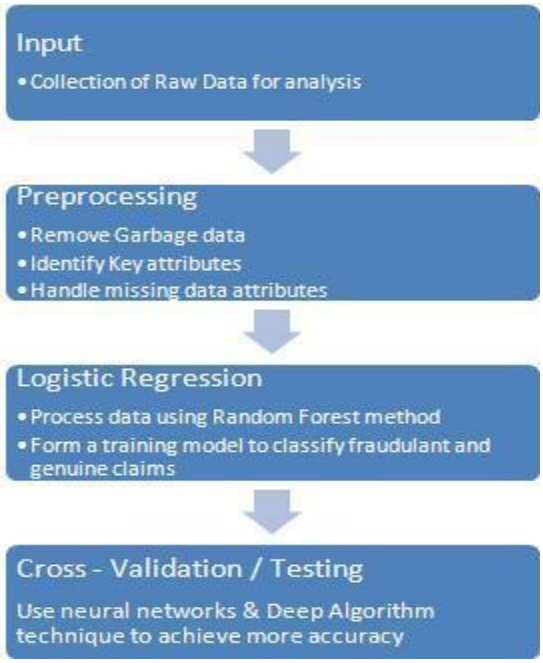

Figure 1. Proposed System

Once the raw data was obtained, the next step was to scrub the data –

a. To fill any missing details
b. To remove unwanted attributes which are not required for processing

The next step will be to form a training model using the 80% of the data which will be used for initial comparison for any new claims to identify fraudulency pattern. Initial classification of the claims would be done using random forest method. After preparing the data, the analysis is performed using various algorithms according to the need.

© 2017, IJSRNSC All Rights Reserved

The method consists of set of tasks to train and create classified and regressive predictive models.

It consists of various steps listed below -

a. Splitting of data
b. Pre-processing of data
c. Selection of important data attributes
d. Tuning the training model (reset any missing numerical values i.e. NA with 0)

The next step would be to apply deep algorithms using neural networks approach on the 20% of data to achieve more accuracy.

The method consist of 3 main parts as mentioned below

a. Input Layer
b. Hidden Layer
c. Output Layer

The hidden layer is where user would be applying deep algorithm and there can be n number of layers in this stage based on the complexity of the claim data to be analyzed. They are termed as hidden as they are not visible as network output. Using neural networks user were able to classify another 8% and achieve overall classification of 88%

B. Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file for "MSW A4 format".

C. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

V. CLASSIFICATION AND REGRESSION

In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node

A. Properties of Random forest

- Intrinsically multi-class
- Handles Apple and Orange features
- Robustness to outliers
- Works w/ "small" learning set
- Scalability (large learning set)
- Prediction accuracy
- Parameter tuning

VI. ALGORITHM FOR BOTH CLASSIFICATION AND REGRESSION

1. Draw n_{tree} bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an un-pruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when m_{try} = p, the number of predictors.)
3. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

An estimate of the error rate can be obtained, based on the training data, by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample using the tree grown with the bootstrap sample.
2. Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate.

VII. RESULT ANALYASIS

As per the experimental result analysis some sample training datasets are tested in the system. It gives the excellent response. Accuracy graph distinguishes between the linear regression and random forest algorithm.

TABLE I. ACCURACY ANALYSIS

Algorithm/Training-Testing DataSize	Percentages		
	70%-30%	80%-20%	90%-10%
Linear Regression (Existing System)	78.35	80.90	87.7
Random Forest	79.89	83.6	88.23

© 2017, IJSRNSC All Rights Reserved

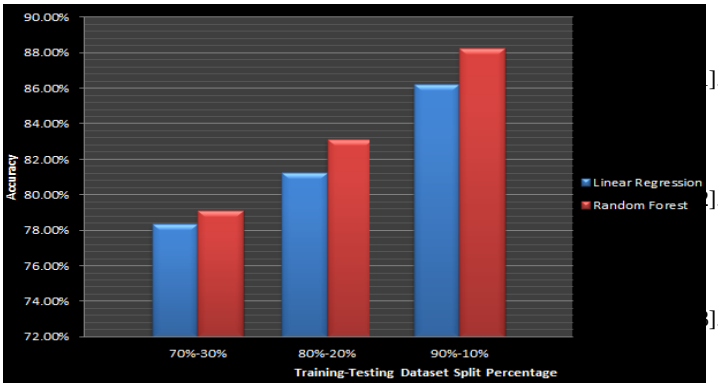

Figure 2. Accuracy Graph

Accuracy graph shows the increased accuracy percentages of the random forest algorithm.

There are some samples results of Transactions Analysis are as follow.

TABLE II. FRAUDULENT TRANSACTIONS ANALYSIS

count	492.000000
mean	80746.806911
std	47835.365138
min	406.000000
25	41241.500000
50	75568.500000
75	128483.000000
max	170348.000000

TABLE III. NON-FRAUDULENT TRANSACTIONS ANALYSIS

count	284315.000000
mean	94838.202258
std	47484.015786
min	0.000000
25	54230.000000
50	84711.000000
75	139333.000000
max	172792.000000

VIII. CONCLUSION

The implemented system detects the fraud in the healthcare or health insurance area. As fraud becomes more sophisticated and the volume of data grows, it becomes more difficult to recognize fraud from bulk of data. Developer may not eliminate fraud but surely reduce it. Data mining uncovers patterns hidden in data to deliver knowledge. Here user have used random forest algorithm and logistics regression algorithm techniques for fraud detection in health insurance. These techniques are more efficient and secure as compare to previous one.

REFERENCES

[1]. Fuzail Misarwala, KausarMukadam, and Kiran Bhowmick, "*Applications of Data Mining in Fraud Detection*", International Journal of Computer Sciences and Engineering, Vol.3, Issue.11, pp.45-53, 2015.

[2]. Min Nelofar Kureshi, Syed Sibte Raza Abidi, "*A Predictive Model for Personalized Therapeutic Interventions in Non-small Cell Lung Cancer*", IEEE Journal of Health Informatics Vol. 20, No.1, pp.424-431, 2016.

[3]. Vipula Rawte, G Anuradha, "*Fraud Detection in Health Insurance using Data Mining Techniques*", International conference of ICCTCT, Mumbai, pp.66-71, 2015.

[4]. Melih Kirlidoga, Cuneyt Asuk(2012) "*A fraud detection approach with data mining in health insurance"*, Procedia Social and Behavioral Sciences, US, pp.989-994, 2012.

[5]. Dan Ventura, " *SVM Example*", BYU University of Physics and Mathematical Sciences, India, pp.1-25, 2009.

[6]. Shunzhi Zhu, Yan Wang, Yun Wu, "*Health Care Fraud Detection Using Nonnegative Matrix Factorization*", The 6th International Conference on Computer Science and Education (ICCSE 2011), Singapore, pp.1-6, 2011.

[7]. Zhongyuan Zhang, Tao Li, Chris Ding, Xiangsun Zhang, "*Binary Matrix Factorization with Applications*", Proceeding ICDM '07 Proceedings of the 2007 Seventh IEEE International Conference on Data Mining Pages, China, pp. 391-400, 2007.

[8]. Mohammad Sajjad Ghaemi, "*Clustering and Nonnegative Matrix Factorization*", Computer Science and Software Engineering Department- Laval University, Victor, pp.1-12, 2013.

[9]. Haesun Park, "*Nonnegative Matrix Factorization for Clustering*", School of Computational Science and Engineering Georgia Institute of Technology Atlanta, USA, pp.1-36, 2012.

[10]. 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, Vol.7, Issue.8, pp.140-144, 2013.

[11]. Williams, G., Huang, Z., "*Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases*", Proc. of the 10th Australian Joint Conference on Artificial Intelligence, Australia, pp.34-44, 1997.

[12]. Wong, W., Moore, A., Cooper, G. ,Wagner, M., "*Bayesian Network Anomaly Pattern Detection for Detecting Disease Outbreaks*", Proc. Of ICML, UK, pp.217-223, 2013.

[13]. Anand, D. Khots, "*A data mining framework for identifying claim overpayments for the health insurance industry*", INFORMS Workshop on Data Mining and Health Informatics", India, pp.47-53, 2008.

[14]. L. J. Opit, "*The cost of health care and health insurance in Australia: Some problems associated with the fee for-service*", Soc. Sci. Med., Vol.18, No.11, pp.967-972, 1984.

[15]. J. Major, D. Riedinger, "*EFD: A hybrid knowledge/statistical based system for the detection of fraud*", Journal of Risk and Insurance, Vol.69, Issue.3, pp. 309-324, 2002.

© 2017, IJSRNSC All Rights Reserved

[16]. V. Krishnaiah, G. Narsimha, N.S. Chandra, "*Survey of Classification Techniques in Data Mining*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.9, pp.65-74, 2014.

[17]. Divya Tomar, Sonali Agarwal, "*A survey on Data Mining approaches for Healthcare*", International Journal of Bio-Science and Bio-Technology, Vol.5, Issue.5, pp.241-266, 2013.

Authors Profile

Prof. Namrata D. Ghuse pursed Master of Engineering from Prof. Ram Meghe Institute of Engineering & Management, Badnera, Amravati in 2014. She is currently working as Assistant Professor in Department of Computer Engineering of Sandip Institute of Technology & Research Centre, Nashik. She is a member of ISTE i.e. International Society for Technology in Education. She has published more than 10 research papers in reputed international journals and it's also available online.

© 2017, IJSRNSC All Rights Reserved