HEALTH CARE INSURANCE FRAUD DETECTION: A DATA MINING PERSPECTIVE

Nikita Borse[1], Neeta Maitre[2]
[1]Student, Computer Department ,Cummins College of Engineering for Women, Pune, India
2Assistant Professor, Computer Department, Cummins College of Engineering for Women, Pune, India
Email: [1]nikita.borse@cumminscollege.in, [2]neeta.maitre@cumminscollege.in

Abstract— Statistics have shown that millions of dollars used for the healthcare expenditures annually are exhausted due to frauds. Within the healthcare sector , Data mining technology plays a vital role in detection of frauds in insurance claims. This research paper gives a brief insight about data mining, information of healthcare insurance frauds and highlights of the advantages of data mining technique (Bayesian Classification) over other methods used in fraud detection.

Keywords—Data Mining, Health care Insurance, Service Provider's Fraud Detection , Bayesian Classification.

I. INTRODUCTION

Health care systems play a important role in the quality of life and in social welfare of the modern society. One of the important sectors in the healthcare systems is the healthcare insurance sector. A significant change in the geometry of healthcare insurance coverage has been observed with the launch of several insurance schemes. But the increasing number of frauds in these schemes are leading to the exhaustion of healthcare expenses. A tremendous volume of data generated by insurance transactions are complex to be processed and analysed by traditional methods.

Data mining is a technology that is used to transform the mounds of data into useful information and helps in decision making that benefits in detecting frauds in the healthcare insurance.

II. HEALTHCARE INSURANCE FRAUD

Healthcare insurance fraud is an intention to distort relevant information for the benefits of an individual or a group of people.[1] About 19-20% of healthcare expenditure is wasted due to fraud and abuse. Therefore, the scale of healthcare fraud is large enough to make it a priority issue for the healthcare systems.

There are 3 parties involved in the healthcare insurance fraud. They are as follows-

1. *Service Providers :* Includes the party who provide health care services to the insurance subscribers and in turn get the payments from insurance subscribers for the services provided. Service providers include doctors, hospitals and laboratories.
2. *Insurance Carriers* : Includes the ones who receive regular premiums from their subscribers and pay them back the settlement amount. Insurance Carriers include governmental healthcare departments and private insurance companies.

3. *Insurance Subscribers* : They render the services from the service providers and make the payments for those services. The subscribers in turn get the reimbursement (settlement amount) from the the insurance carriers. Insurance Subscribers include patients and patients' employers.

According to which party commits the fraud, fraud behaviours can be classified as follows-

1) Service Provider's Fraud include:

- Billing of services which are not actually rendered.
- Unbundling – Billing each stage of services as if tit were a separate treatment.
- Upcoding – Billing more costly services than actually performed.
- Performing unnecessary medical services for generating more payments.
- Falsifying patient's diagnosis or the treatment history to justify tests, surgeries that are not medically necessary.

2) Insurance subscriber's fraud include:

- Falsifying the records of employment for obtaining a l lower premium rate.
- Filing claims for healthcare services which are not actually received.
- Using the coverage of some other person to illegally obtain the healthcare benefits.

3) Insurance Carrier's Fraud include:

- Falsifying the reimbursements.
- Falsifying the benefits/service statements. [1] The main focus in this research paper is on detection of the above mentioned Service Provider's fraud using data mining technology.

III. DATA MINING

Data mining is one of the vital areas of research, for finding meaningful information from large datasets. Data Mining is a core of the **Knowledge Discovery Process (KDD) .** [2] The steps involved in the KDD process are • Data Cleaning - Preprocessing of data.

- Data Integration – Multiple data sources are combined. • Data Selection – Data relevant for analysis is retrieved from databases.
- Data Transformation – Data is transformed and consolidated by performing aggregation operations to get appropriate data for mining.
- Data Mining – It is an essential process where different methods are applied to extract meaningful data patterns.
- Pattern Evaluation – Identification of truly interesting patterns representing knowledge.
- Knowledge Representation – Visualizations techniques ` used to present mined knowledge to users. [6] Data mining is thus considered to be one of the important steps of KDD process.

A. Classification of Data Mining :

Data mining can be classified according to:

- Kind of the data being mined.
- Kind of the knowledge being discovered.
- Kind of the techniques / algorithms.

The most common and well accepted categorization is :

1) Unsupervised Techniques :

This technique is used when no prior information is available i.e. no training data set is available. In this method. Clusters are used when there is no historical data available. Clusters are basically the similar kind of data patterns grouped together. Examples include – Association Rules, Clustering and Outlier (Anomaly) Detection.

2) Supervised Techniques:

This technique is used when training dataset is available. It is useful in analysing new pattern based on the previously known patterns and discovering the relationship between input and output variables. It is faster in terms of implementation as compared to Unsupervised methods.

Examples include - statistical methods like Regression Analysis, Neural Networks, Bayes Classification and
Support Vector Machine. [6]

B. Applications of Data mining :

Data mining has applications in several industries :

- Transportation
- Banking
- Quality Control
- Healthcare
- Finance
- Telecommunication

The focus here is data mining applications in Healthcare.

C. Applications of Data Mining in Healthcare Sector

Data mining offers the healthcare sector the capabilities to tackle eminent challenges germane to its domain. Among them being :

1) Evaluation of Treatment Effectiveness :

Using Data mining, physicians can discover the patient's health and lifestyle histories as they can impact medical coverage and service utilization. Data mining also helps them to identify the side effects of particular treatment, to make appropriate decision to reduce the hazard and to develop smart methodologies for treatment.

2) Customer Relationship Management :

Data mining helps the healthcare institute to understand the needs, preferences, behaviour, patterns and quality of their customer in order to make better relation with them

3) Detection of Fraud in Insurance Claims:

Healthcare insurer develops a model to detect the fraud in insurance claims using data mining techniques. This model is helpful in identifying the potential fraud. This is applicable to measure improper prescriptions, irregular or fake patterns in claims made by physicians, patients, hospitals etc. [4]

Statistics have shown that millions and crores of rupees of the governmental organization and insurance companies are exhausted in frauds. Several organizations are unable to detect the frauds using traditional methods. In such cases, data mining appears as a useful technology helping in minimizing exhaustion of the expenditures.
In this way, data mining is beneficial in several applications especially in the insurance fraud detection within the healthcare sector.

IV. DATA MINING TECHNIQUES FOR DETECTING FRAUD

A) Review of Available Studies :

1) SAS Enterprise Miner (4.2) :

Commercial software by SAS and CLUTO (Software developed by University of Minnesota) has used data mining technique called Clustering technique. The observational results are that – CLUTO takes less computational time as compared to SAS. CLUTO takes 6.5 minutes whereas SAS takes several hours.[5]

2) The Health Insurance Commission in Australia used an online unsupervised learning algorithm called as 'Smartsifter' to detect outliers in the pathology utilization services in Medicare Australia. It was basically used to detect the anomalies within the services provided. [2]

3) Liou et al (2008) has used supervised methods to review insurance claims of diabetic outpatient services that were submitted to Taiwan's National Health Insurance. A comparative study on three data mining techniques – logistic regressions, neural networks and classification tress was done for the detection of fraudulent claims and they came to a conclusion that all the three methods gave accurate results but the Classification model performed the best with the overall identification rate of 99% [2]

B) Comparative Studies :

Classification method is found to be more accurate as compared to the other methods according to the studies. The Classification methods are categorized as Decision Tree, Naive Bayesian Classifier, Rule based classification and so on.

The following table shows a comparative study of two classification algorithms- Naive Bayesian and Decision Tree providing the advantages of Naive Bayesian Classification over Decision Tree.

TABLE I
COMPARATIVE STUDY OF DECISION TREE AND NAIVE BAYESIAN CLASSIFIER

Decision Tree	Naive Bayesian Classifier
Decision Tree is a tree based classifier.	Naive Bayesian is a probability classifier.
It does not assume that the attributes are independent.	Assumes the variables or attributes to be independent.

Larger the decision tree, less accurate will be the results.	Highly accurate results on large datasets.

V. IMPLEMENTATION : BAYESIAN CLASSIFICATION

To classify whether the new study of insurance claim is fraudulent or not, we can use one of the classification methods of data mining. **Naive Bayes Classifier** is one of the classification techniques that provides high accuracy and speed on large datasets.

D. Naive Bayesian Classifier:

It is a probability classifier and it predicts the class membership probabilities such as the probability that a given tuple belongs to a particular class or not. It assumes that an attribute value on a given data is independent of the other attribute values.

Such an assumption is called a Class Conditional Independence. As this assumption is made for reducing the comparative costs, hence called as Naive.[3]

Naive Bayes Classifier is based on the Bayes Theorem. According to this theorem "X" is a data tuple which is considered as the evidence and is described by the set of n attributes.

This algorithm considers that the data entries are classified into classes. These classes are decided prior by the user itself. "H" is the hypothesis such as a data tuple (X) belongs to the specified class C. The Bayes Theorem is stated as follows -

$$P(H|X) = P(H)\ .\ \frac{P(X|H)}{P(X)}$$

Here, P(X|H)- Posterior Probability, P{X|H)- Likelihood, P(H)- Class Prior Probability , P(X)- Prior Probability of X.

E. Training Data Set :

The data attributes for the training data set were taken from The United States Health Information Knowledge-base.

As the Service Provider's frauds have to be detected, therefore the combination of attributes are selected according to it. For simplicity, only few of these attributes are selected for creating a smaller training dataset.

F. Test Data Set :

Four attributes are considered while providing the test data as follows - Actual Services, Standard Services, Claimed Amount , Settlement Amount.

The test data with the values are passed to check whether the entry is fraudulent or not. If the probability indicates that it is a fraud then, the inference is that service providers had manipulated the actual count of services that were to be provided to the patients and if the output comes as not fraud, then the inference would be that the test data was not fraudulent and no service provider's fraud is detected.

In this way, we can detect whether the test data is fraudulent or not using the Naives Bayesian Probability
Classifier.

VI.CONCLUSIONS

Service Provider's fraud detection is only one small part of the bigger program of combating the healthcare insurance fraud.

The traditional methods of fraud detection are time consuming and inefficient, hence Data mining is a better methodology as it provides fast and accurate results . Data mining includes several techniques, out of which Naive Bayes Classification is one such technique, which provides high accuracy and speed over the large datasets.

Thus, Naive Bayes classifier is a better approach to detect such types of frauds (Service Provider's Fraud), which will help in minimizing the loss of the healthcare expenditures which are lost due to the frauds.

REFERENCES

[1] Arash Rashidian, Behrouz Minaei-Bidgoli, Bijan Geraili, Hos-sein Joudaki, Mahdi Nasiri, Mahmood Mahmoodi & Mohammad Arab, "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature", *Global Journal of Health Science*, Vol. 7 , Issue. 1, 2014.

[2] Divya Tomar, Sonali Agarwal, " A survey on Data Mining approaches for Healthcare", *International Journal of Bio-*

Science and Bio-Technology, Vol. 5, Issue. 5, 2013, pp 241-266. [3] K.Kalaiselvi , R. Bhuvaneswari, “Naive Bayesian Classification Approach in Healthcare Apllications”, *International Journal of Computer Science and Telecommunication,* Vol. 3, Issue. 1,2012, pp 106-112.

[4] Jianjun Shi, Jing Li, Jionghua Jin, Kuei-Ying Huang, “A survey on statistical methods for health care fraud detection”, *Health Care Management Science,* Vol. 11, Issue. 3 ,2008, pp 275-287.

[5] Alan Sabatka, Deepak Khazanchi, Gang Kou, Yi Peng, Yong Shi, Zhengxin Chen, “Application of Clustering Methods to Health Insurance Fraud Detection”’, *IEEE*, Vol .1,October,2006. [6] Jian Pei, Jiawei Han, Micheline Kamber, “Data Mining :Con cepts and Techniques”, 2004, *The Morgan Kaufmann Series*, 3rd ed.