Health Insurance Claim Fraud Detection: A Survey

V. Kathiresan
Department of Computer Science and Engineering
CIET, Coimbatore, Tamilnadu , India

Dr. S. Gunasekaran
Department of Computer Science and Engineering
CIET, Coimbatore, Tamilnadu , India

Faseela V. S
Department of Computer Science and Engineering
CIET, Coimbatore, Tamilnadu , India

Abstract- The anomaly or outlier detection takes vital role in data mining. The outlier detection techniques again play an important role in credit card transaction fraud detection, health insurance claim fraud detection and other web usage fraud detections. A huge exigency is there for effective fraud detection mechanism to improvise the health insurance management system. This paper exhibit and analyze the performance of various techniques used in different paper to detect health insurance claim fraud in comparative approach. The Spectral analysis, Support vector machine and Multilayer neural network (MLP) are the techniques which is analyzed and surveyed in a comparative manner in this paper. The effectiveness and role of each technique in fraud detection is examined and the pros and cons are reviewed in this paper.

Keywords – Spectral Analysis, Support Vector Machine, Neural Networks.

I. INTRODUCTION

In several countries fraudulent behavior in health insurance claim is a major problem. Data mining tools and techniques can be used to detect fraud in large sets of insurance claim data. One of the most common data mining techniques used to find fraudulent records is anomaly detection. This paper aims to review various approaches used for Health insurance claim fraud detection. There are three major parties involved in the entire system,

(1) Service Providers
(2) Insurance Subscribers
(3) Insurance Carriers

The Service Providers are doctors, hospitals, ambulance companies and laboratories. The Insurance Subscribers are patients and patient's employers. The Insurance Carriers who receive regular premiums from subscribers and pay health care cost on behalf of their subscribers. There is a difference between fraud prevention and fraud detection. The fraud prevention describes measures to avoid fraud to occur. The fraud detection involves identifying fraud as quickly as possible once it has been committed.

According to the National Health care anti-fraud association, health care fraud is the misrepresentation of Claims for gaining some shabby benefits. The health industry in India is losing approximately Rs.600 crores on "false claims" every year. So to make health insurance feasible, there is a need to focus on eliminating or reducing fraudulent claims. There are two types of frauds, first one is Hard fraud, This is a deliberate attempt either to point an event or an accident, which requires hospitalization or other type of loss that would be covered under a medical insurance policy. Second one is Soft fraud, which occur when people purposely provide false information such as claim fraud, application fraud and eligibility fraud sources and then put to use by data miners to achieve the desired results.
The rest of the paper is organized as follows. Some of the techniques for health insurance management are explained in section II. Comparative study and Comparative graph are presented in section III and IV respectively. Concluding remarks are given in section V.

II. HEALTHCARE FRAUD DETECTION TECHNIQUES

Data Mining for Healthcare Management is an emerging potential area with respect to its impact on improving healthcare as a result of discovering new patterns and trends in voluminous data generated by healthcare transactions. Some of the existing approaches of data mining for health insurance fraud management are been listed below,

A. Using Spectral Analysis --

In this approach[2], the health care claim datasets are passed in to Nodes list and Edges list. The Nodes list consists of two sets of Nodes, one is a list of all the primary care physicians(PCPs) and another is a list of all the specialists. The Edge list is a list of edges that connect all the PCPs and Specialists.

A two-mode network is constructed using the Nodes list and Edges list. This network is transformed in to a Laplacian matrix, L(u,t). The eigenvalues and eigenvectors are computed for L(u,t) using the fielder vector or the eigenvector of the second smallest eigenvalues of L(u,t),Thus obtained the connectivity feature of this network.

From the real world healthcare dataset, first create a node list and an edge list. A list of 1,101 PCPs and a list of 7,823 Specialists form the two-mode network. There are a total of 8,834 nodes and 60,207 edges connecting between these nodes. The nodes can be denoted as,

$$\text{Node (i)} = \begin{cases} PCP & if\ i \geq 7824, i \in [1,8834] \\ Specialist & if\ i < 7824, i \in [1,8834] \end{cases} \quad (1)$$

This paper[2] proposing Gap cut algorithm in which first determine the minimum gap between values in order to separate different communities. It is simple and efficient in the spectral analysis to divide the nodes in to communities of unknown numbers. The goal is to detect the suspicious communities between PCPs and specialists in health care claim datasets.

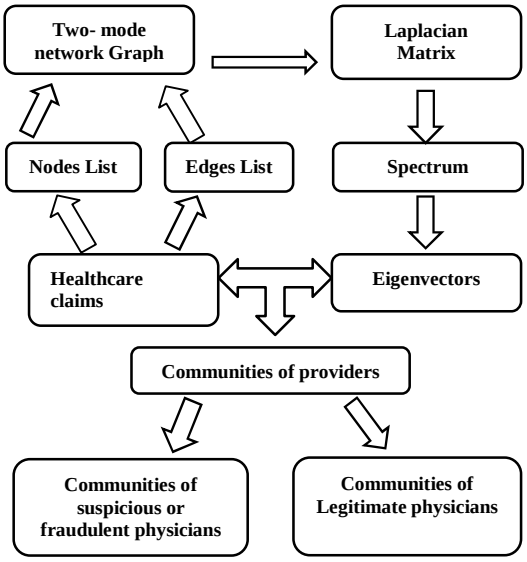

Figure 1. Overall Approach

B. Using Support Vector Machines –

This paper[4] focuses on workers compensation bills, which are often audited more aggressively than group health bills. In particular, types of abuse that involve the misuse of codes, for example, unbundling is a type of abuse in which a procedure that is supposed to be charged as a unit is broken down in to its component procedures which can often increase the amount that can be charged for a bill. Such abuse is difficult to detect except by a trained medical expert.

By performing data cleansing and scalability, bills get standardized and elementary errors resolved. For the learning tasks, the implementation of linear SVM is used. The Linear SVM is well regarded for scaling of both number of features and observations.

(i) Encoding

Many critical features are medical codes. As features, represent the codes with a sparse binary encoding. The binary parity allows both training and testing speed to be in elastic under the addition of further binary features.

(ii) Feature Selection

The features, which is determined to be most used are such as occurrences of codes, whether ICD9, CPT or modifier codes appear in a bill, tax identification number of the billing hospital, number of lines in the bill, number of days of service, total duration of the bill.

(iii) Adjudication Type

The type of facility is called adjudication type. It is depend on whether a bill is from an inpatient hospital, an ambulance, an emergency room etc. There are different kinds of expected or possible procedures and practices and different acceptable payment practices. Before analysis can be performed on a bill, its adjudication type must be determined so that an analyst can reason about reasonable payments. From table 1, an average of 99% accuracy with a minimum of 98% across all of the adjudication types can be obtained[4].

Table -1 Adjudication classification confusion matrix

Predicted	ASC	PRO	AMB	IPH
ASC	99.5	.27	0	0
PRO	.36	99.21	0	0
AMB	0	0	99	0
IPH	1.5	0	.25	98

(iv) Abuse Detection

Abusive bills are classified with an unacceptably high false positive rate[4]. From table 2, it is easier to predict impactedness than abuse. Abusive bills are classified with an unacceptably high false positive rate.

Table -2 Abuse classification confusion matrix

Predicted	Abuse	Not Abuse
Abuse	72%	28%
Not Abuse	7%	93%

(v) Minimizing False Positives

The bagging schemes are useful to overcome the high false positive rate. There are two parameters, number of bags and number of bills per bag. These parameters have to be optimized and have found two optima, one more accurate and the other minimizing false negatives. The first optima are found when bagging fifty models each containing 2/3 of the original training set. With these parameters, accuracy was highest and received a significant decrease in false negative rate[4].

Table- 3 Abuse classification Results with 50 2/3 of Training set bags and 90% Threshold

predicted	Abuse	Not Abuse	On-hold
Abuse	79%	19%	2%
Not Abuse	5%	94%	1%

C. *Using Multilayer Neural Networks* --

This paper[5] describes the fraud or abuse detection system that utilizes one committee of multilayer neural networks (MLP) for each one of the entities involved in the fraud or abuse problem (medical claims, affiliates, medical professionals &employers). This divides and conquers strategy allows to feedback information over time, combining affiliates, doctors and employer's behavior.

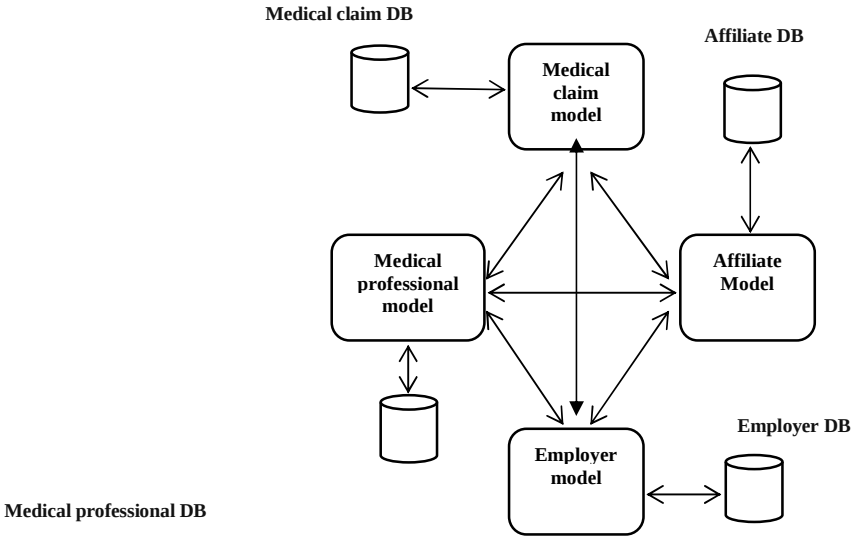

Figure 2. Fraud detection scheme by using sub-models with feedback connections

(i) Entities and Medical Claim Data

A medical claim involves the participation of an affiliate, a medical professional, and an employer. The data such as age, sex, type of claim, affiliate's name and date of birth, ID number, resting period solicited, type and place of the resting, identification of the medical professional, identification of the employer, labor activity of the company where the affiliate works, affiliate's profession and income records.

(ii) Business and Data Understanding

Several meetings were held with some medical experts who explained the main aspects of criteria for approval, modification and rejection of medical claims. This allowed better understanding the underlying business model, including discriminative behavioural patterns, as well as weaknesses of the current non standardized fraud detection procedure. The outcome of these meetings was a preliminary set of variables designed to discriminate between normal and fraudulent behavior.

(iv) Data Preparation

To build a robust classifier using only a small training set, it was decided to apply a divide and conquer strategy. The initial problem was subdivided in to smaller problems namely, four separate models to cope with the entities. Each sub model required a smaller number of features and training samples. First, an exhaustive manual classification was done, assisted by both medical experts and legal advisors. Second, the feature vectors for the four different models were further optimized. Finally standard data preparation techniques have been applied to avoid training biases.

(v) Modeling

After the medical claim submission, different entities are analyzed separately using historical data with cross references among them. Initially, each sub problem was modeled by a standard two layer neural network (MLP). The classifier's accuracy showed a high variance that exceeded 8% between different runs. In order to decrease the model variance, a committee of multilayer feed forward neural networks is implemented. Each sub-model had committees of 10 MLPs. Table 4 compares the standard deviation obtained for single network against a committee[5].

Table- 4 Employer Model, Classification variance

Classifier	Mean	Std. Dev.
Single MLP	88.7%	8.4%
Committee(10MLPs)	88.7%	1.8%

(vi) Incorporation in to Fraud Detection Workflow

The digitalized medical claim forms arrive during each day. The automated fraud detection system is executed at night, assigning a fraud probability to each form. Associated affiliates, medical professionals and employers are updated as well. At the next day, a web interface allows to consult these records ranked by their fraud probabilities, acting like a prescreen filter. The historical data considering a total of 8819 employers was analyzed. This set contains 418 fraudulent/abusive and 8401 normal cases. This dataset was divided in to a training, validation and test set[5].

Table -5 Employer Dataset

Employer category	Number of classes	percentage
Fraud/abuse (T)	176	20.0
Normal (T)	706	80.0
Total (T)	882	100.0
Fraud/abuse (V)	118	20.1
Normal (V)	470	79.9
Total (V)	588	100.0
Fraud/abuse (T)	124	1.7
Normal (T)	7,225	98.3
Total (T)	7,349	100.0

The illustration of efficacy of this fraud detection system shows the evolution of an employer's fraud score. The fraud score evolution is given in Table 6. This fraud detection scheme identifies fraud/abuse employer behavior within 2 months, whereas the former procedure took 8.6 months in the average[5].

Table -6 Fraud score evaluation, employer id 7771xxx

Fraud/abuse probability	Reception date of medical claim
0.6118	September 13,2005
0.6908	October 13,2005
0.8246	October 18,2005
0.9126	October 31,2005

This first record originated from an affiliate's medical claim submission. The employer sub-model delivered a value 0.61 of being fraudulent or abusive. Because there is no history associated to the employer until that point. The second medical claim arrived on October 13, 2005. After the update, employer sub-model scored this record with an abuse probability of 0.69. The third medical claim arrived on October 18, 2005 and fourth on October 31. After the update, employer sub-model score augmented to 0.82 and 0.91 respectively. The experts confirmed the abuse case and activated an alarm rule that highlights future medical claims from this employer. The divide and conquer approach is able analyze each involved entity separately, along with other beneficial side-effects such as dimensional reduction and the consequent model robustness.

III. COMPARATIVE STUDY

Table- 7 Comparative Study

TITLE	TECHNIQUE USED	MERITS	DEMERITS
A Novel Approach to Uncover Health Care Frauds Through Spectral Analysis	**Spectral Analysis**	**Detect the suspicious communities between PCPs and Specialists in health care claim datasets**	**Performs fraud detection in Two mode network only**
Using Support Vector Machines to Detect Medical Fraud and Abuse	**Support Vector Machines(SVM)**	**Level of accuracy obtained is acceptable**	**Human auditors are needed**
A Medical Claim Fraud/Abuse Detection System based on Data mining: A Case Study in Chile	**Multilayer Neural Networks(MLP)**	**Uses divide and conquer approach**	**Estimation of cost and savings cannot be performed**

IV. COMPARATIVE GRAPH

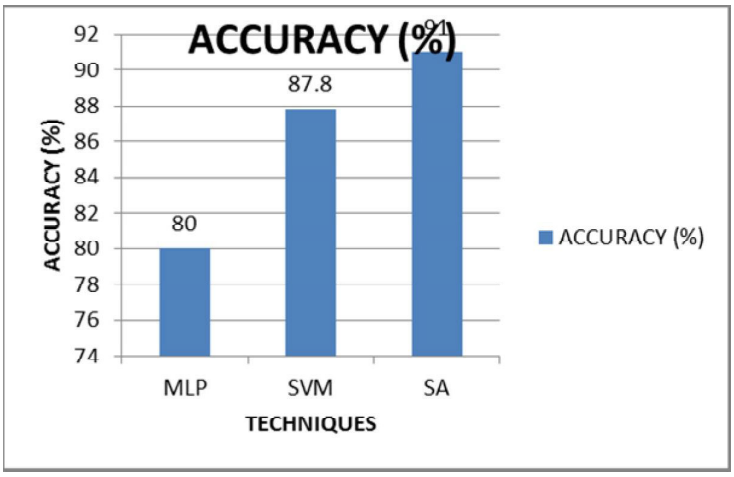

Figure 3. Comparative graph

MLP – Multi Layer Neural Networks SVM - Support Vector Machine SA -Spectral Analysis

The above graph depicts the accuracy analysis of various techniques used for health insurance claim fraud detection. The spectral analysis method shows highest accuracy than other two techniques. The accuracies are calculated by using this formula,

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

V. CONCLUSION

In conclusion, this paper reviews various approaches for detecting fraudulent behavior in health insurance claim. By analyzing the aforementioned techniques, we will get a clear idea for the future work in health insurance claim fraud detection. In India, we have three levels of health care network, namely primary, secondary, and tertiary. It provides an opportunity for data miners to use the huge amount of data. The main task is to integrate data from different sources and then put to use by data miners to achieve the desired results.

REFERENCES

[1] Ah Chung Tsoi, Shu Zhang, Markus Hagen Buchner, “Hierarchical Hidden Markov Models an Application to Health Insurance Data”, 2006.
[2] Song Chen , Aryya Gangopadhyay, ”A Novel approach to Uncover Health Care Frauds Through Spectral Analysis”IEEE International Conference on Healthcare Informatics,2013.
[3] Shunzhi Zhu, Yan Wang, Yun Wu,”Health care Fraud Detection Using Nonnegative Matrix Factorization “The 6th International Conference on Computer Science & Education, 2011.
[4] Charles Francis, Noah Pepper, Homer Strong, ”Using Support Vector Machines to Detect Medical Fraud and Abuse”33rd Annual International Conference of the IEEE EMBS Boston,USA,2011.
[5] Pedro A.Ortega, Cristian J. Figueroa, Gonzalo A. Ruz,”A Medical Claim Fraud/Abuse Detection System based on Data mining: A Case Study in Chile “International Conference on Data mining, USA, 2006.
[6] Qi Liu, Miklos Vasarhelyi,”Health fraud detection: A survey and a clustering model incorporating Geo-location information”2013.
[7] Melih Kirlidog , Cuneyt Asuk ,”A fraud detection approach with data mining in health insurance”,2012.

[8] Dallas Thornton, Roland M. Mueller, Paulus Schoutsen, Jos van Hillegersberg, ”Predicting Healthcare Fraud in Medicaid :A Multidimensional Data Model and Analysis Techniques for Fraud Detection”,2013.
[9] H.Lookmansithic, T.Balasubramanian,”Survey of Insurance Fraud Detection Using Data Mining Techniques”, International Journal of innovative Technology and Exploring Engineering, 2013.
[10] Anand Sharma, Vibhakar Mansotra,”Emerging applications of Data mining for Health care Management-A Critical Review” IEEE,2014.
[11] Jiawei Han, Micheline kamber, Jian pie, “DATA MINING Concepts and Techniques” Third Edition, 2012.