Fraud Detection in Health Insurance using Hybrid System

Thotakura Lalithagayatri
Student
Department of Computer Engineering
Atharva College of Engineering, Mumbai,
Maharashtra, India

Tawde Priyanka
Student
Department of Computer Engineering
Atharva College of Engineering, Mumbai,
Maharashtra, India

Aruna Pavate
Assistant Professor
Department of Computer Engineering
Atharva College of Engineering, Mumbai,
Maharashtra, India

***Abstract*— Health claim frauds are affecting the economic status of developing as well as developed countries. Health care fraud detection is now becoming more and more important. In order to detect and avoid fraud we are going to use data mining techniques. We have proposed a Hybrid model system consisting of classification and clustering. Considering all the advantages and disadvantages of algorithm involved in classification and clustering, Evolving clustering method and Support vector machine are chosen. The Fraud claims will be detected and the genuine claims will be paid by the insurance company.**

Keywords*— Hybrid system; SVM; ECM; Healthcare fraud; insurance claims;

I. INTRODUCTION

Data mining techniques are applied to detect and avoid frauds. This includes some Preliminary knowledge of health care system and its fraudulent behavior, analysis of the characteristics of health care insurance data. Fraud is widespread and very costly to the health care insurance system. The main purpose of fraud is financial benefit. The Aim of our project is to detect fraud in health insurances.

Frauds blow a hole in the insurance industry. The National Health Care Anti-Fraud Association (NHCAA) estimates that the losses due to health claims fraud are in the tens of billions of dollars every year [6]. Health insurance is a sector with very high claims ratio. So, to make health insurance industry free from fraud, it is necessary to eliminate or minimization of fake claims arriving through health insurance. The development of a safe, high-quality, and cost-effective health care system requires effective ways to detect fraud. This system may not eliminate fraud but can surely reduce it.

The health insurance fraud claims are broadly classified as:

• Billing for services not rendered: Claiming insurance for services that never happened. Example: Fake bills.

• Up coding of services: Claiming insurance for services that are costlier than the original. Example: Admitted for 3 days and claiming insurance for 5 days.

• Up coding of items: Claiming insurance for items are costlier than the actual items. Example: Medicines of 1000 Rupees were claimed as 5000 Rupees.

• Duplicate claims: Some changes are made in original bill and claimed to the insurance company again for second time.

•Unnecessary services: Claiming insurance company for unnecessary services. Example: A non-cancer patient claims for chemotherapy.

The main purpose of fraud is financial benefit. According to a recent survey, it is estimated that the 15 per cent of total claims are fraud. Insurance companies in USA incur losses over 30 billion USD annually to healthcare insurance frauds. The statistics is appalling in developing country like India as well. According to the Healthcare industry in India is losing approximately Rs. 600-Rs 800 crores incurred on fraudulent claims annually. Frauds blow a hole in the ratio. So insurance industry. Health insurance is a sector with very high claims, to make health insurance industry free from fraud, it is necessary to focus on elimination or minimization of fake claims arriving through health insurance [1].

II. LITERATURE SURVEY:

Vipula Rawte had proposed the individual implementation of Evolving Clustering Method and Support Vector Machine. They had chosen Evolving Clustering Method (ECM) for clustering because the data is dynamic i.e. the claims are dynamic and new data is generated continuously and Support Vector Machine (SVM) is used for classification [1]. Rashmi Dutta Baruah, Plamen Angelov and Diganta Baruah had implemented the clustering system. In that system the evolving clustering approach attempts to meet the following three key requirements of data stream clustering: (i) fast and memory efficient (ii) adaptive (iii) robust to noise. But the disadvantage of this system is that Clusters of new diseases are formed but duplicates are not detected [2].

Sriram Ravindran, Chandan Gautam, Aruna Tiwari have implemented Extreme Learning Machine and Evolving Clustering Method. The problem of recognizing a user from the passphrase is performed using ELM and ECM-ELM. Stable accuracies were obtained from ECM-ELM. Accuracies were good but need to be improved [3]. Lijuan Liu, Bo Shen, Xing Wang3 had introduced the theoretical basis of support

vector machine, summarize the research status and analyze the research direction and development prospects of kernel function. The kernel function is used in Support Vector Machine to resolve the errors occurred during classification of datasets [4]. Janmenjoy Nayak, Bighnaraj Naik* and H. S. Behera had performed a survey on Support Vector Machine. The main aim of this paper is to deduce the various areas of SVM with a basis of understanding the technique and a comprehensive survey, while offering researchers a modernized picture of the depth and breadth in both the theory and applications [5].

Jing Li & Kuei-Ying Huang & Jionghua Jin & Jianjun Shi, 2007- had done research on fraud healthcare claims. This paper is the first to provide a comprehensive survey of published research results in health care fraud detection. They made efforts to classify the fraudulent behaviors, identify the sources and characteristics of health care data, provide key Steps in data preprocessing, and summarize and compare existing statistical methods [6].

III. PROPOSED SYSTEM

There are two sections in existing system:

A. CLASSIFICATION*:* It can classify claims as fraudulent or genuine. Whenever a new unknown disease comes, the classification algorithm classifies it as fraud even if it is not. Only primitive diseases are identified & classified if new one is detected it's directly classified as fraudulent [1].

B. CLUSTERING: It can deal with new unknown disease but it cannot classify claims as fraudulent or genuine. Clusters of new diseases are formed but duplicates are not detected. [1]

Major drawback of supervised and unsupervised techniques are that the supervised technique cannot classify claims of an unknown disease while the clustering technique cannot detect outliers when duplicate claims i.e. claims with different dates are filed.

There are many Algorithms available for Classification and Clustering. Following are the drawbacks of different algorithms for classification and clustering:

- C4.5 ALGORITHM: Small variation M data can lead to differ decision trees does not work very well on a Small training data set.
- ID3 ALGORITHM: Requires large searching time. Requires large amount of memory to store tree
- NAIVE BAYES ALGORITHM: Prediction rate decreases on small data. For good results it requires large number of records.
- ARTIFICIAL NEURAL NETWORK ALGORITHM: Requires high processing time. Difficult to know how many layers are required.
- K-NEAREST NEIGHBOUR ALGORITHM: Time to find the nearest neighbours in a large training data set can be excessive.
- SUPPORT VECTOR ALGORITHM: Speed and size requirement is more. High complexity and extensive memory required [4].

So to overcome the drawbacks of such classification and clustering algorithm we have proposed the hybrid model for detecting health insurance frauds and flag them for further investigation. In Hybrid System we have chosen Evolving Clustering Method (ECM) for clustering because the data is dynamic i.e. the claims are dynamic and new data is generated continuously and Support Vector Machine (SVM) is used for classification. In this approach, first, the insurance claims are clustered according to the disease type by ECM and then they are classified to detect any duplicate claims by SVM.

Hybrid System:

Insurance claims are clustered by applying the ECM algorithm and then these clusters are given to SVM algorithm for classification. As a result, clusters get formed for all the diseases' claims including the new unknown disease which won't be possible with traditional clustering method like k-means clustering technique. So, cluster gets formed for Parkinson's disease claims as well. Next, the duplicate claim won't get detected on applying clustering. This drawback is overcome by applying classification based on date on the already formed clusters.

IV. METHODOLOGY

Insurance claims are clustered by applying the ECM algorithm and then these clusters are given to SVM algorithm for classification. As a result, clusters get formed for all the diseases' claims including the new unknown disease which won't be possible with traditional clustering method like k-means clustering technique. So, cluster gets formed for Parkinson's disease claims as well. Next, the duplicate claim won't get detected on applying clustering. This drawback is overcome by applying classification based on date on the already formed clusters. Hence, SVM classifies the duplicate claim. Thus, the hybrid approach of ECM and SVM shall prove to be useful in medical health insurance domain for detecting the health insurance frauds.

C. EVOLVING CLUSTERING METHOD:

ECM is used to cluster dynamic data. Dynamic data are those which keep on changing with respect to time. As and when new data point comes in, ECM clusters them by modifying the position and size of the cluster. There is a parameter known as radius associated with each cluster that determines the boundaries of that cluster. Initially, the cluster radius is set to zero. The radius of the cluster increases as more data points are added to that cluster. It has one more parameter known as the distance threshold Dthr, which determines the addition of clusters. If the threshold value is small then, there will be more number of small clusters and if the value is large, then there will less number of large clusters. Selection of the threshold is dependent on the heuristics of the data points.

D. SUPPORT VECTOR MACHINE:

A support vector machine is a supervised learning technique used in classification. It has an initial training phase where data that has already been classified is fed to the algorithm. After the training phase is finished, SVM can predict into which class the new incoming data will fall into.

SVM Steps:

- Training (Preprocessing Step):

1. Define two class labels viz. "legitimate" or "fraudulent"
2. Classify claims into two classes using the training data set.
3. Choose support vectors and find the maximum marginal hyper plane that separates the claims into two classes.

- Classification:

Identify the new incoming claims into either "Legitimate" or "fraudulent" class.

E. Steps in Hybrid Model Construction:

• Doctor bills patients for the services/equipment given to them during their treatment.

• Patient files claims to the insurance company.

• Claims are submitted to the Hybrid Framework wherein clustering (ECM) is followed by classification (SVM) to detect the fraudulent claims.

• There is an expert who flags the fraudulent claims for further investigation with the insurance company.

• The legitimate claims are further passed to the insurance Company and those claims are paid to the patients

F. Pseudo Code for the Hybrid Approach:

• Apply ECM to each of the incoming health insurance claim to form clusters according to the disease type.

• Apply SVM to each of the clusters to classify those claims into "legitimate" and "fraudulent" classes.

Go back to clustering step to cluster new claims and repeat.

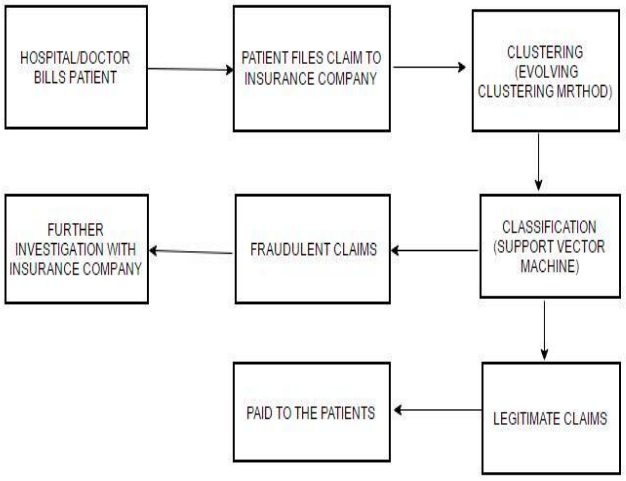

Fig.1 Control flow of system

V. CONCLUSION AND FUTURE WORK

The disastrous effect of health care fraud needs to be reduced. The propose Hybrid system may not completely eliminate fraud but surely reduce it. Data mining involves mainly classification and clustering techniques. Considering the advantages and disadvantages of most of the classification and clustering techniques, ECM and SVM are Chosen. Evolving clustering method clustering method can cluster dynamic data, hence it is chosen. The clustered data is then classified but Support Vector Machine. In Future proposed Hybrid System can be implemented.

REFERENCES

[1] Vipula Rawte, "Fraud Detection in Health Insurance using Data Mining Techniques" International Conference on Communication, Information & Computing Technology (ICCICT),Mumbai, India, Jan. 16-17,2015.

[2] Rashmi Dutta Baruah, Plamen Angelov and Diganta Baruah,"Dynamically Evolving Clustering for Data Streams",2014

[3] Sriram Ravindran, Chandan Gautam, Aruna Tiwari,"Keystroke User Recognition through Extreme Learning Machine and Evolving Cluster Method",IEEE International Conference on Computational Intelligence and Computing Research,2015.

[4] Lijuan Liu,Bo Shen, Xing Wang3,"Research on Kernel Function of Support Vector Machine"

[5] Janmenjoy Nayak, Bighnaraj Naik* and H. S. Behera,"A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges", International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.169-186 http://dx.doi.org/10.14257/ijdta.2015.8.1.18, 2015.

[6] Jing Li & Kuei-Ying Huang & Jionghua Jin & Jianjun Shi, "A survey on statistical methods for health care fraud detection", 2007.