Fraud Detection in Health Insurance using Data Mining

Ms. Pavitra Chandrakant Nhavkar[1] Prof. Anushree Gaud[2]
[1,2]Department of Computer Application
[1,2]Bharati Vidyapeeth's Institute of Management & Information Technology, India

Abstract— Fraud detection is a technique of identifying forbidden acts. Because of the problem of frauds, there is a need of various fraud detection techniques to detect the frauds. This will help the auditors and management to detect previously undetected frauds also and develop various controls to prevent such frauds from reoccurring in the future.
Key words: Fraud Detection, Health Insurance, Data Mining

I. INTRODUCTION

Claiming money from insurance company using illegitimate approach by an individual is known as health insurance fraud. Such illegitimate claiming of insurance by individuals or groups has caused the insurance companies to face a lot of losses. The frauds in health insurance are increasing every year. A survey done in USA revealed that the false claims in the health insurance industry is approximately 15 percent of the total. This has caused the health insurance industry to face a loss of about 30 billion USD annually [6]. In order to reduce the losses incurred because of these frauds data mining techniques are widely used to detect them. This includes knowledge about health care system and behavior of frauds in them. Data mining concepts are widely used for analyzing data patterns.

II. LITERATURE SURVEY

Some of the chosen Supervised and Unsupervised data are as follows:

A. Anomaly Detection

Anomaly detection also known as outlier detection can be defined as the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.[8] It detects some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.[9]

B. Support Vector Machines

Support Vector Machines (SVM) is a powerful, state-of-the-art algorithm with strong theoretical foundations based on the Vapnik- Chervonenkis theory. SVM has strong regularization properties. Regularization refers to the generalization of the model to new data.[5]
Fraud claims in health insurance industry are classified as follows:

1) Upcoding of Services

Charging more than the actual/expected cost of the procedure performed on a patient.

2) Identical Claim

In order to get the benefits twice the individual or group submits another copy of the same bill by manipulating some data in the bill.

3) Unnecessary Services

Mentioning claims which doesn't apply to the patient's condition.

4) Charging for Services not taken by the Patient

This happens when the individual or group claims for services which were not rendered by him/her.

5) Up Coding of Items

Manipulating with the actual cost of the items used for treating patient in order to gain more money.

III. DATA MINING

Data mining is when we discover new patterns which are unknown before and reliable statistically. With the use of data mining we can understand data patterns with huge datasets. The various data mining techniques are Clustering, Classification, Advanced neural networks, Prediction and regression models.[3]
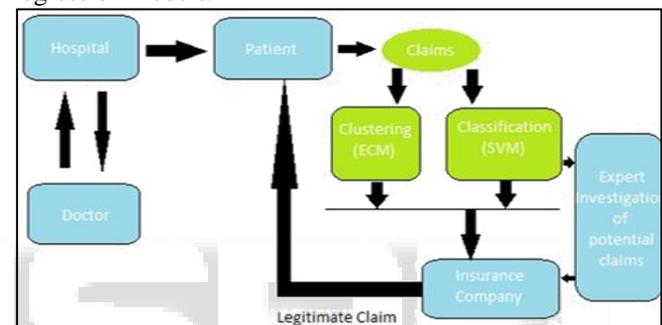

Fig. 1:

Fraud detection is the identification of frauds. First we have to find that whether given data is fraudulent or not. Supervised learning on datasets is used to deal with fraud data that is previously known and unsupervised learning on datasets issued to deal with the fraud data that was previously considered as a legitimate data. But after processing the same data shows the characteristics of fraud data. Hence by merging these two approaches i.e supervised and unsupervised an innovative approach for identifying fraud claims in health insurance industry is proposed.
We will be considering following supervised and unsupervised data mining techniques:

A. Support Vector Machines

Support Vector Machines work on decisions taken on the basis of boundary values. The system identifies and segregate the data in two types legitimate and fraud data. Once the whole processing is done every single claim is compared with the boundary. Based on the results obtained the claims are divided into legitimate or fraud.

B. Anomaly Detection

In this technique the probability is calculated on each claim to check if it is fraud or not. A background check of previous claim is also done.

C. Non Negative Matrix Factorization

Every medical item used for treating a patient is clustered differently for each individual patient. These clusters are then grouped based on the symptoms of a particular disease. The

fraud claim are detected by examining the shift in the clusters for one medical item compared to the other items and their respective clusters.[4]

D. *K-Means Algorithm*

It takes parameter k as input and further divides n objects into k clusters. This results in high intra cluster similarity and low inter cluster similarity. The number of clusters are predefined in this algorithm[4].These fixed number of clusters also becomes the drawback for this system as it becomes difficult for new incoming objects.

E. *Outlier Detection*

In this technique a default value is used for consumption of a medicine for a particular disease. If the resultant values are deviated from the default value shows the presence of fraud claim. The analysis is done on the basis of clustering[5].

IV. POSSIBLE SOLUTIONS FOR DETECTING FRAUDS

The supervised and unsupervised technique has a few limitations respectively, like the supervised technique cannot be used to check if the claim is valid or not when the disease is not known. Similarly the unsupervised technique cannot find claim with manipulated data i.e claims with different dates. To eliminate this limitation, two methods can be used Evolving Clustering Method and Support Vector Machine. The ECM method is used for handling dynamic clustering as new data is constantly introduced to the system. And the SVM technique is used for classification of the claims. According to this approach the data about the insurance claims are first clustered based on different type of disease. Once the claims are clustered they are further classified as legitimate of fraud[7].

V. BLOCK DIAGRAM

Above figure shows the model for fraud detection considering ECM and SVM

A. *Steps Followed*

− Hospital charges patients through bills.
− Patient submit the bills to insurance company for claiming.
− Clustering and Classification of the claims using the ECM and SVM technique is performed on the claims using respective techniques.
− Experts identify potential fraud claims based on the results of ECM and SVM and send the results to the insurance company.
− Insurance company then further passes the claim based on the results given by the expert i.e the legitimate patients are paid with the claim amount by the insurance company.

VI. CONCLUSION

Fraud have increased over the past few years. It is also becoming more and more complex to handle large datasets. These large datasets introduced new challenges for detecting fraud claims. To counter these challenges we are using data mining techniques such as clustering and classification. Data mining can efficiently help in identifying hidden patterns in

data and to extract knowledge which can be eventually used for identifying fraud claims done by an individual or a group. In this paper we have described a method which used ECM and SVM for clustering and classification of data respectively. ECM is used as it support dynamic data and SVM provides scalability which provides good quality of data.

REFERENCES

[1] Guido Cornelis van Capelleveen, "Outlier based Predictors for Health Insurance Fraud Detection within U.S. Medicaid", University of Twente & University of California, San Diego December 2013.
[2] https://www.researchgate.net/publication/282538462_Fraud_detection_in_health_insurance_using_data_mining_techniques
[3] Dan Ventura. Class Lecture, Topic: "SVM Example." BYU University of Physics and Mathematical Sciences, Mar. 12, 2009.
[4] Dan Ventura. Class Lecture, Topic: "SVM Example." BYU University of Physics and Mathematical Sciences, Mar. 12, 2009.
[5] Dan Ventura. Class Lecture, Topic: "SVM Example." BYU University of Physics and Mathematical Sciences, Mar. 12, 2009.
[6] Dan Ventura. Class Lecture, Topic: "SVM Example." BYU University of Physics and Mathematical Sciences, Mar. 12, 2009.
[7] Haesun Park. Class Lecture, Topic: "Nonnegative Matrix Factorization for Clustering". School of Computational Science and Engineering Georgia Institute of Technology Atlanta, GA, USA, July 2012.
[8] 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
[9] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_svm.htm#DMCON021
[10] Leonard Wafula Wakoli. "APPLICATION OF THE K-MEANS CLUSTERING ALGORITHM IN MEDICAL CLAIMS FRAUD/ABUSE DETECTION." MSc Thesis, Jomo Kenyatta University of Agriculture and Technology, 2012.
[11] Mohammad Sajjad Ghaemi. Class Lecture, Topic: "Clustering and Nonnegative Matrix Factorization". DAMAS LAB, Computer Science and Software Engineering Department, Laval University. Apr.12, 2013.
[12] Chandola, V.; Banerjee, A.; Kumar, V. (2009). "Anomaly detection: A survey". ACM Computing Surveys. 41 (3): 1–58.
[13] Hodge, V. J.; Austin, J. (2004). "A Survey of Outlier Detection Methodologies" (PDF). Artificial Intelligence Review. 22 (2): 85–126.