Fraud Detection in Health Insurance using Data Mining Techniques

Vipula Rawte
Student, M.E. (Computer Engineering)
St.Francis Institute of Technology
Mumbai-400103, India.
rawtevipula25@gmail.com

G Anuradha
Associate Professor (Computer Engineering)
St.Francis Institute of Technology
Mumbai-400103, India.
ganusrinu4@yahoo.co.in

***Abstract*—Fraud is widespread and very costly to the health-care insurance system. 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. In order to detect and avoid the fraud, data mining techniques are applied. This includes some preliminary knowledge of health care system and its fraudulent behaviors, analysis of the characteristics of health care insurance data. Data mining which is divided into two learning techniques viz., supervised and unsupervised is employed to detect fraudulent claims. But, since each of the above techniques has its own set of advantages and disadvantages, by combining the advantages of both the techniques, a novel hybrid approach for detecting fraudulent claims in health insurance industry is proposed.**

Keywords—data mining; health insurance fraud; supervised; unsupervised

I. INTRODUCTION

Deliberately deceiving the health insurance company that results in healthcare benefits being paid illegitimately to an individual or group is known as health insurance fraud. The main purpose of fraud is financial benefit. According to a recent survey, it is estimated that the number of false claims in the industry is approximately 15 per cent of total claims. 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. The report suggests that the healthcare industry in India is losing approximately Rs 600-Rs 800 crores incurred on fraudulent claims annually [1]. Frauds blow a hole in the insurance industry. Health insurance is a bleeding sector with very high claims ratio. So, to make health insurance industry free from fraud, it is necessary to focus on elimination or minimization of fake claims arriving through health insurance.

The health insurance fraud claims are broadly classified under the following headings:

- 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.
- 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. DATA MINING

Nowadays there is huge amount of data stored in real-world databases and this amount continues to grow fast. So, there is a need for semi-automatic methods that discover the hidden knowledge in such database. Data mining automatically filtering through immense amounts of data to find known/unknown patterns, bring out valuable new perceptions and make predictions.

Data mining techniques tend to learn models from data. There are two approaches on learning the data mining models. Those are supervised learning, unsupervised learning; and they are described below:

A. Supervised Learning:

This is the most usual learning technique wherein the model is trained using pre-defined class labels. In the context of health insurance fraud detection the class labels may be the "legitimate" and "fraudulent" claims. The training dataset can

978-1-4799-5522-0/15/$31.00 ©2015 IEEE

be used to build the model. Then any new claim can be compared with the already trained model to predict its class. A claim will be classified as a legitimate claim if it follows a similar pattern to the legitimate behavior else it will be classified as an illegitimate.

The advantages are that all classes are meaningful to humans and it can be easily used for pattern classification.

The disadvantage is the difficulty associated in gathering class labels. Also, when there is bulk input data, it is costly to label all of them, and claims must be identified properly because false positives and true negatives can create a bad impression about the insurance company in the minds of its customers. Skewed distribution – of the class labels in the training dataset can result in a model which does not have a very good accuracy for prediction. Supervised learning models cannot detect new types of frauds and significant efforts are required from the experts to derive the labeled training samples which will be used to construct the model.

B. Unsupervised Learning:

Unsupervised learning has no class labels. It focuses on finding those instances which show unusual behaviour. Unsupervised learning techniques can discover both old and new types of fraud since they are not restricted to the fraud patterns which already have pre-defined class labels like supervised learning techniques do.

The advantages are it aims to detect anything which does not abide by the normal behaviour and because of the lack of direction, it can find patterns that have not been noticed previously.

While the disadvantage being because of lack of direction, there may be times when no interesting knowledge has been discovered in the set of features selected for the training.

III. LITERATURE REVIEW

There are many supervised and unsupervised data mining techniques out of which the following are chosen.

ANOMALY DETECTION: The anomaly detection technique calculates the probability of each claim to be fraudulent by examining the previous insurance claims. The analysts further investigate the cases that have been flagged by data mining model [2].

SUPPORT VECTOR MACHINES: SVM is fundamentally a classification technique. The system is trained to determine a decision boundary between classes of “legitimate” and “fraudulent” claims. Then each claim is compared with that decision boundary and is placed into either legitimate or fraudulent class [2], [3].

NON-NEGATIVE MATRIX FACTORIZATION: It is a technique for clustering medical treatment items into several clusters according to usage by different patients. Each cluster can be shown as group of medical treatment items for curing symptoms of similar diseases. Now, this technique can identify fraud if a medical treatment item shifts from one cluster to another in a period of one month. Its drawback being it is intractable to solve. [4],[5],[6],[7].

k-MEANS ALGORITHM: The k-means algorithm takes the parameter k as input, and divides a set of n objects into k clusters such that the resulting intra-cluster similarity is high while the inter-cluster similarity is low. This algorithm predefines the number of clusters. This becomes the drawback for clustering new incoming objects since there would be fixed number of clusters [8],[9].

OUTLIER DETECTION: Here, a baseline of the ususal behaviour of usgae of medicne dosgae for patient is established. Any deviation from this baseline indicates an outlier. It generally results from clustering [10].

Supervised techniques (classification) cannot classify new types of disease claims whereas unsupervised techniques (clustering) cannot detect duplicate claims i.e. fraud. So, a novel hybrid approach for health insurance fraud detection is proposed. The drawbacks of supervised and unsupervised techniques are explained with the help of examples below.

A. Advantage of Supervised Technique (Classification) over Unsupervised Technique (Clustering):

Consider there are two claims made by the same patient, out of which one is the original claim and the other one is a duplicate claim as shown in Fig. 1. Duplicate claim is formed by changing the date but keeping rest of the patient’s details same.

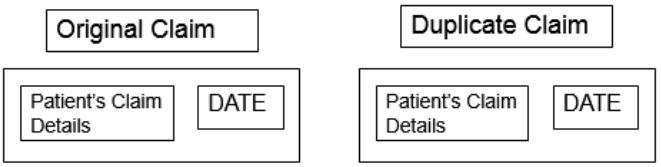

Fig. 1. Original and duplicate claims made by the same patient

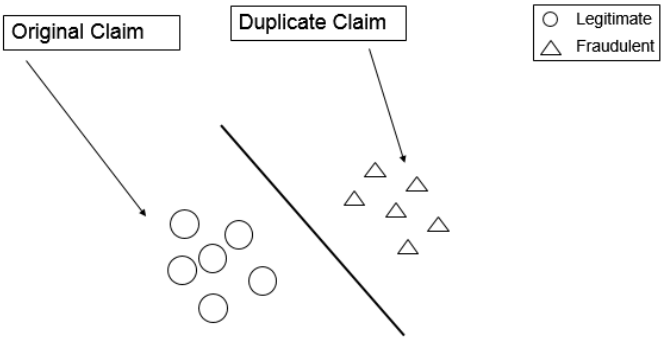

Fig. 2. Classifying the claims

In Fig. 2, both the claims (original and duplicate) get classified into their respective classes based on the training given to the SVM. Here, the duplicate claim gets classified into the fraudulent class and hence gets detected.

978-1-4799-5522-0/15/$31.00 ©2015 IEEE

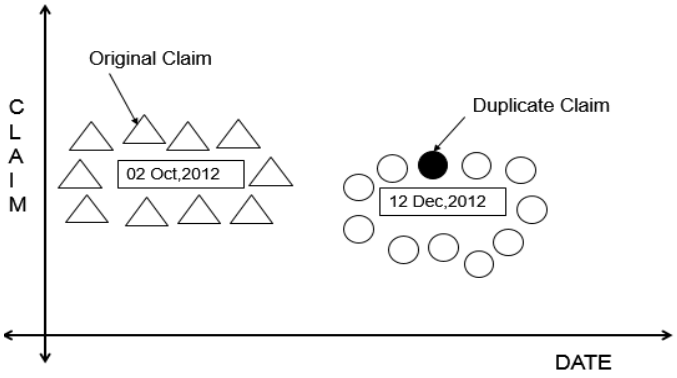

Fig. 3. Classification (SVM) succeeds over Clustering (Outlier)

But, Fig. 3. shows that if clustering based approach like outlier detection is used, then the duplicate claim doesn't get identified.

B. Advantage of Unsupervised Technique (Clustering) over Supervised Technique (Classification):

Consider that the SVM can classify only the following types of diseases' medical claims.

- Heart Disease
- Arthritis
- Dyslexia
- Diabetes
- Cancer
- Kidney Failure
- Paralysis
- Alzheimer's Disease

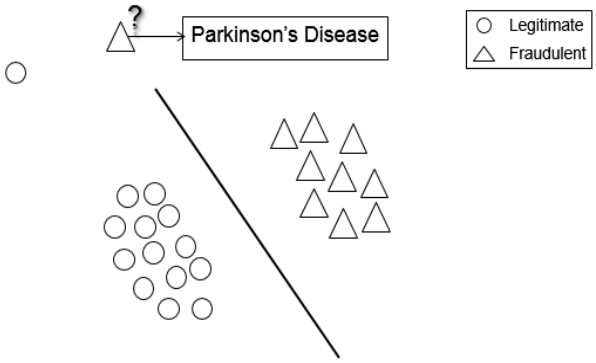

Fig. 4. Claasification (SVM) unable to classify Parkinson's Disease Claim

In Fig. 4., SVM could not classify some new disease like Parkinson's Disease claim.

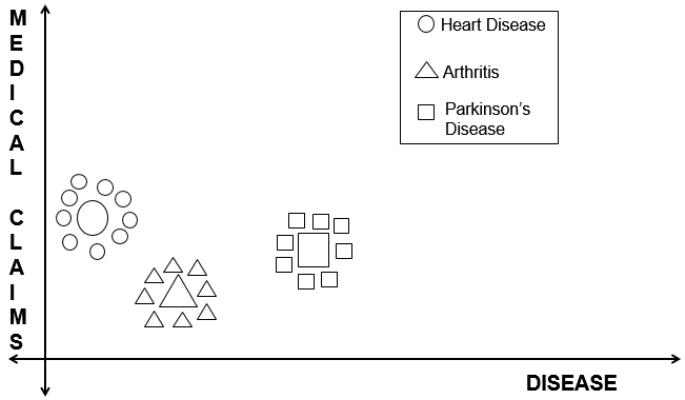

Fig. 5. Clustering groups different claims according to the diseases' type

Fig. 5, shows that clustering could identify and cluster the Parkinson's Disease claims.

It is clear from the above example, that both, classifications as well as clustering have their own set of advantages and disadvantages. So, it would be better if advantages of both are combined together to form a new hybrid approach. Such a novel approach is proposed in the next chapter.

IV. Proposed Approach

Major drawback of supervised and unsupervised techniques are that the former cannot classify claims of an unknown disease while the latter cant detect outliers when duplicate claims i.e. claims with different dates are filed. So, in this section, we propose a hybrid model for detecting health insurance frauds and flag them for further investigation. For this, we have chosen Evolving Clustering Method (ECM) for clustering because the data is dynamic and new data is generated continuously and Support Vector Machine (SVM) for classification. In this approach, first, the insurance claims are clustered according to the disease type and then they are classified to detect any duplicate claims. So, ECM and SVM are explained in the following sections.

A. Evolving Clustering Method (ECM):

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 D_{thr}, which determines the addition of clusters [11]. 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. Fig. 6 shows the flowchart of Evolving Clustering Method (ECM).

978-1-4799-5522-0/15/$31.00 ©2015 IEEE

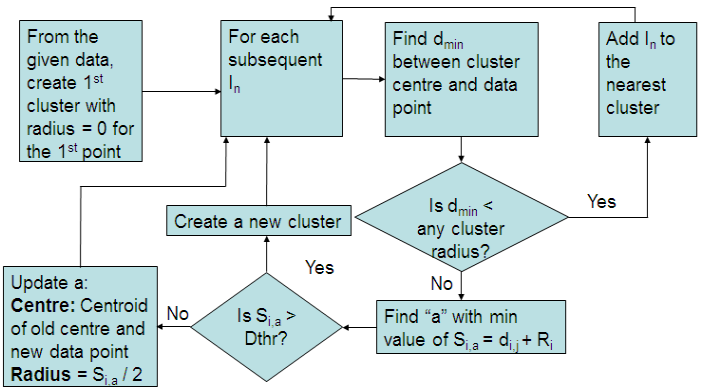

Fig. 6. Flowchart for ECM

ECM – Example:

Consider the following table with the given data points:

TABLE I. ECM EXAMPLE

x co-ordinate	y co-ordinate
1	1
1.1	1
1	1.1
2	3
6	6
1.1	1.1
6.1	6
6	6.1

Let us assume the threshold = 0.2

Four clusters viz., A, B, C, D are created which are shown in Fig. 7.

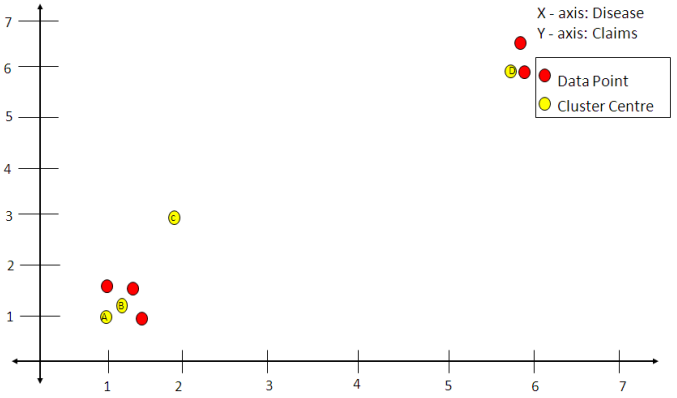

Fig. 7. ECM Example

B. Support Vector Machine (SVM):

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:

1) Training (Preprocessing Step):

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

2) Classification:

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

C. Block Diagram:

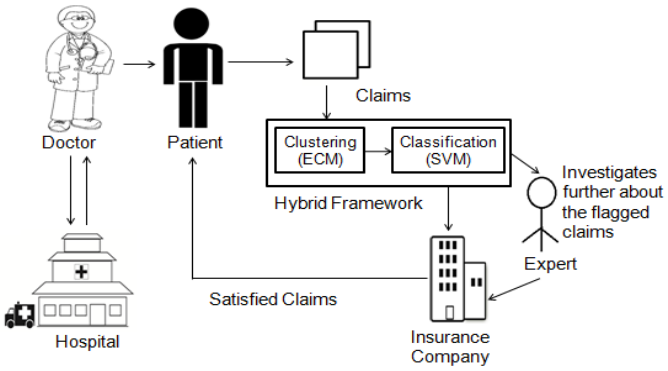

Fig. 8. Hybrid Model

Considering ECM and SVM, Fig. 8 shows the block diagram for the hybrid model of fraud detection.

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.

D. Pseudo Code for the Hybrid Approach:

- For each of the incoming health insurance claim, apply ECM 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

978-1-4799-5522-0/15/$31.00 ©2015 IEEE

Consider an example where there are duplicate claims of the unknown (new) disease.

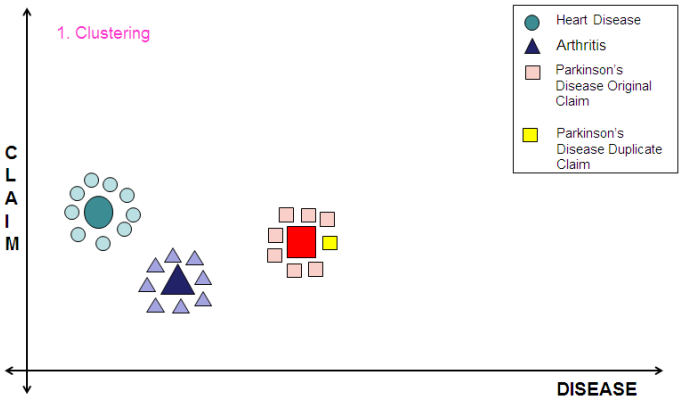

Fig. 9. Clusters are formed according to disese type

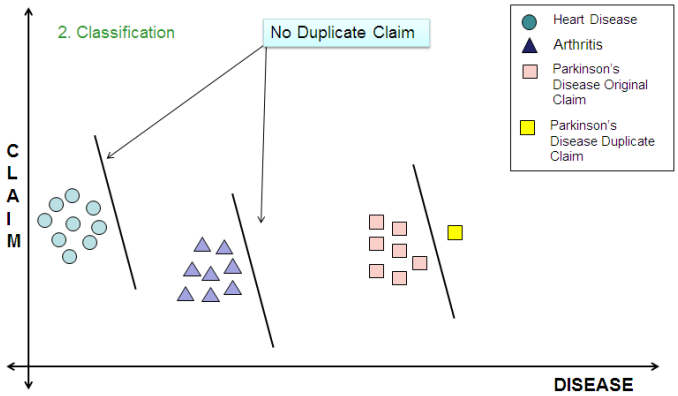

Fig. 10. Duplicate claim is detected by classifying it as a fraudulent claim

From Fig. 9 and Fig. 10, it is clear that first the health 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.

V. Conclusion

As fraud becomes more sophisticated and the volume of data grows, it becomes more difficult to recognize fraud from bulk of data. We may not eliminate fraud but we can surely reduce it. Data mining uncovers patterns hidden in data to deliver knowledge. Data mining involves mainly classification and clustering techniques. Considering the advantages and disadvantages of most of the classification and clustering techniques, ECM is chosen as the clustering technique because the data flows in continuously and there is a need to cluster dynamic data and SVM as classification technique since it provides the scalability and usability that are needed in a good quality data mining system and the quality of generalization and ease of training of SVM is far beyond the capacities of traditional methods such as neural networks and radial basis functions.

References:

[1] Dr.Biswendu Bardhan. "Frauds in Health Insurance", http://healthcare.financialexpress.com/200711/market13.shtml.

[2] Melih Kirlidoga, Cuneyt Asuk(2012) A fraud detection approach with data mining in health insurance. Procedia - Social and Behavioral Sciences 62 (2012) 989 – 994.

[3] Dan Ventura. Class Lecture, Topic: "SVM Example." BYU University of Physics and Mathematical Sciences, Mar. 12, 2009.

[4] Shunzhi Zhu, Yan Wang, Yun Wu, "Health Care Fraud Detection Using Nonnegative Matrix Factorization", The 6th International Conference on Computer Science & Education (ICCSE 2011) August 3-5, 2011. SuperStar Virgo, Singapore.

[5] 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 391-400.

[6] Mohammad Sajjad Ghaemi. Class Lecture, Topic: "Clustering and Nonnegative Matrix Factorization". DAMAS LAB, Computer Science and Software Engineering Department, Laval University. Apr.12, 2013.

[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] 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.

[10] 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.

[11] Qun Song, Nikola Kasabov, "ECM — A Novel On-line, Evolving Clustering Method and Its Applications", Department of Information Science, University of Otago.

978-1-4799-5522-0/15/$31.00 ©2015 IEEE