Applications of Data Mining in Detecting Fraudulent Health Insurance Claim

Dileep Dayanand Pai, Pramod Agnihotri,
Rajath G R,
Assistant Professor,
Department of Information Science & Engineering,
AMCEC

Bineet Kumar Jha
BE (Information Science & Engineering),
AMC Engineering College, Bangalore.

Abstract:- **Fraud exists in each and every part of the health care system. Fraud involves intentional deception or misrepresentation intended to result in an unauthorized benefit. Survey says that fraud in health insurance is an upward graph. To overcome this problem we use big data techniques like data mining. This includes some primary knowledge of health care system and its fraudulent behaviours, analysis of the characteristics of health care insurance data. To detect fraudulent claims we use two types of data mining techniques viz., supervised and unsupervised. Both the techniques have their own positives and negatives we combine the positives of both the techniques to have a hybrid approach for detecting fraudulent claims in health insurance industry.**

Keywords- Data mining; Health insurance fraud; Supervised; Unsupervised

I. INTRODUCTION

The main purpose of fraud is financial benefit. The health insurance companies are deliberately deceived by illegitimate payment of health insurance to an individual or a group is known as fraud in health insurance. Statistics show that approximately 15 to 20 per cent of the total claims are unlawful. Insurance companies in USA incur losses over 30 billion USD annually to health care insurance frauds. The statistics is hideous in developing countries like India. A study suggeststhat the health care industries in India are losing approximately Rs 600 to Rs 800 crores incurred on fraudulent claims annually. To make health insurance industry free from fraud, the fake claims have to be effectively eliminated or minimized.

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

- Forging of bills: Billing insurance company for things that never occurred. Example: Forging the signature of those involved in giving bills, bribing the medical professionals.
- Upcoding of services: Billing insurance that are costlier than the actual services made.
  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.
- Duplicate claims: Producing forged bills by changing the date, price and product names.
- Unnecessary services: Claiming insurance amount from the insurance company for false diagnostic reports.

The amount of data stored in database, based on the real world information is humongous. Hence, there is a need for semi-automatic methods that fish out the hidden data from the database. Data mining automatically filters through immense amounts of data to find hidden patterns, showcase valuable new perceptions and make prediction.
There are two approaches on learning the data mining models. They are supervised learning and unsupervised learning.

A. Supervised Learning:

Supervised learning technique is data mining task of deducing a function from a labelled training data. In the context of health insurance fraud detection the class labels may termed as “legitimate” and “fraudulent” claims. The training dataset can be used to build the model. This helps in comparing the already trained model with the new claim to predict its class. A claim will be declared legitimate if it follows or has a similar pattern as that of the already trained model, else it will be termed as illegitimate.
The advantages of this type are that, all classes here are useful to humans and can be easily used to classify patterns.

The disadvantage is that it is difficult to fetch the class labels and the huge amount of input data results in increasing the cost to label all the input data. The claims must not consider the false positives and true negatives because that leads to a bad impression on the insurance company in the views of customers. Skewed distribution of the class labels should not be made as they do not predict the results accurately.

B. Unsupervised Learning

Class labels do not exist here, instead it focuses on finding the unusual behaviour. This type of data mining technique can discover both old and new types of fraud as they are not restricted to pre-defined class labels or patterns like supervised learning techniques.

The advantages are, it detects 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. The disadvantage is lack of direction.

II. LITERATURE SURVEY

There is *n* number of data mining techniques in supervised and unsupervised categories. They are:-

ANOMALY DETECTION: This technique examines the previous insurance claims and then calculates the probability of each claim to be fraudulent. Further analysis is made by the analyst who investigates the cases that have been flagged by data mining model [2].

SUPPORT VECTOR MACHINES [SVM]: It is basically a classification technique. The system is trained to determine the decision boundary between the classes of “legitimate” and “fraudulent” claims. Then each claim is compared with the result of the decision boundary and is placed into its respective classes. [2], [3].

NON-NEGATIVEMATRI FACTORIZATION: It is a technique to group medical items into many groups based on the usage of it by different patients. Each cluster can be shown as a group of medical treatment items to cure similar diseases. If there is any fraud in medical treatment items through the shifting of items from one group to another, then it is detected by this technique. The only drawback being is it cannot be tracked to solve. [4], [5], [6], [7].

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

A. Advantage of Supervised Techniques (Classification) over UnsupervisedTechnique (Clustering):

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

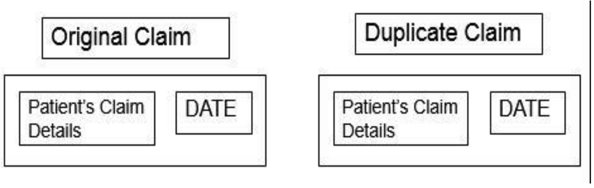

Fig.1. original and duplicate claims made by same patient

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

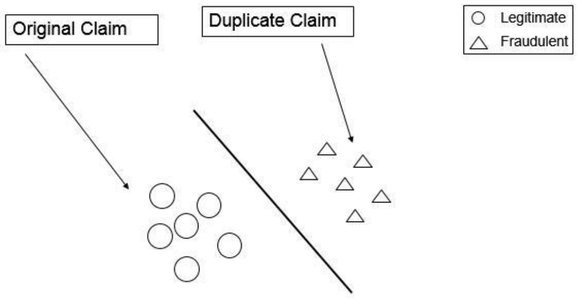

Fig.2. Classifying the claims

Fig. 3, Shows if clustering based approach like outlier detection is used, then duplicate claim doesn’t get identified.

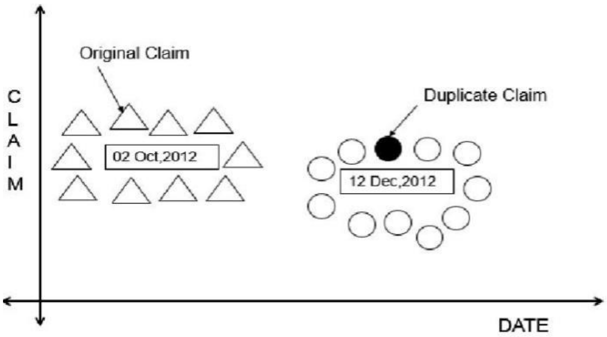

Fig . 3. Classification (SVM) succeeds overClustering (Outliner)

B. Advantage of unsupervised Technique (Clustering) over Supervised Technique (classification):

Taking into consideration that SVM can classify only following types of disease’s in medical claims.

- Heart disease&Arthritis
- Diabetes&Cancer
- Kidney Failure
- Paralysis
- Alzheimer’sDisease

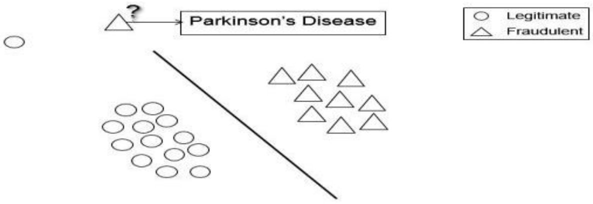

Fig. 4. Classification (SVM) unable to classify Parkinson’s disease claim

In Fig.4. SVM couldn’t classify some claims of new diseases such as Parkinson’s disease.

Well in the next figure (Fig.4.) clustering could identify and cluster the Parkinson’s disease claim.

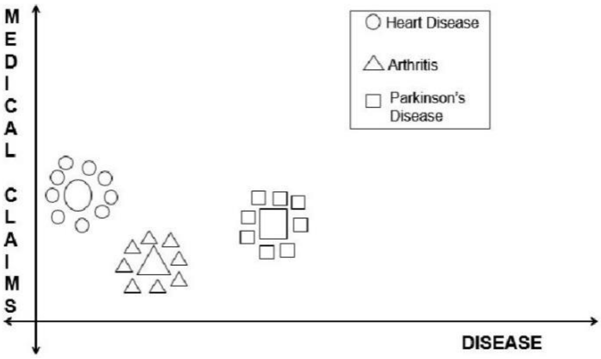

Fig 5. Clustering groups different claims according to disease.

It is clear from Fig 5, that both classification's as well as clustering has its own advantages and disadvantages. So advantages of both are together combined to form a hybrid approach.

III. PROPOSED APPROACH

Major drawback of supervised and unsupervised techniques are former cannot classify claims of an unknown disease while latter cannot detect outliers when duplicate claims. To overcome this new hybrid model proposed to detect frauds in health insurance and flag them for further investigations. A new method for clustering is chosen which is Evolving Clustering Method (ECM) and Support Vector Machine for classification. Here, insurance claims are clustered according to disease type and then they are classified to detect duplicate claims. So SVM and ECM are explained in following section.

A: Evolving Clustering Method (ECM):

ECM is used to cluster dynamic data. Dynamic data keeps on changing with respect to time. As new data comes in ECM clusters them by modifying position and size of the cluster.

There is a parameter associated with each cluster that determines the boundaries of that cluster. Initially radius is set to zero. As radius increase as more as data points are added to cluster it has one more parameter known as distance threshold *Dthr*, which determines the addition of clusters[10]. If threshold is small then more number of small clusters will be there, if large then less number of large clusters. Selection of threshold is dependent on heuristics of data points. Fig.6 shows the flowchart of ECM.

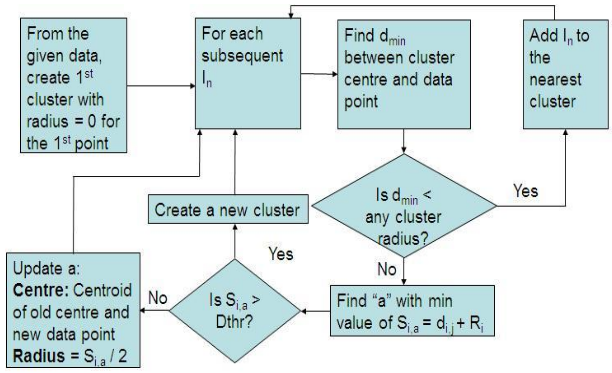
From the given data, create 1st cluster with radius = 0 for the 1st point
For each subsequent I_n
Find d_{min} between cluster centre and data point
Add I_n to the nearest cluster
Is d_{min} < any cluster radius?
Yes
No
Create a new cluster
Find "a" with min value of $S_{i,a} = d_{i,j} + R_i$
Is $S_{i,a}$ > Dthr?
Yes
No
Update a: Centre: Centroid of old centre and new data point Radius = $S_{i,a}$ / 2

Fig.6. Flowchart for ECM

TABLE 1: ECM -EXAMPLE

X co-ordinate	Y co-ordinate
1	1.1
1.1	2
2	3
6	5
1.1	1.1
6.1	6
6	7

Let us assume threshold=0.2

Four clusters viz. A B C D are created as shown in Fig.7. ECM

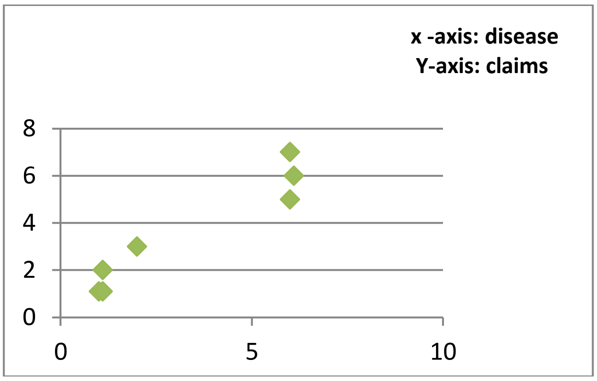

Fig.7. ECM example

B. Support Vector Machine (SVM):

SVM is one of the supervised learning techniques used in classification. The initial phase where data has already been classified is fed to algorithm. After this phase is finished, SVM can predict into which class the mew incoming data will fall into.

SVM Steps:

1) *Training (Pre-processing):*

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

2)*Classification:*

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

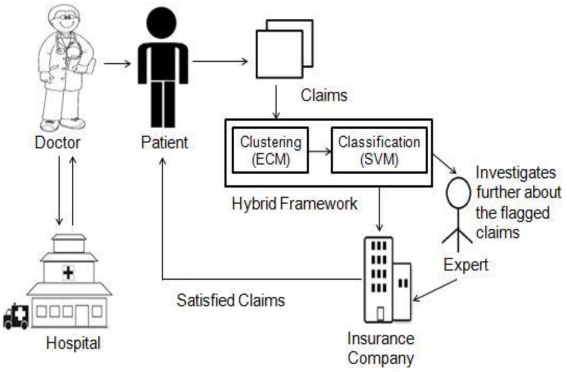

Fig.8. Hybrid model

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

Steps in Hybrid Model consideration:

- Doctor bills patients for the services/equipment given to them during their treatment.
- Patient files claim to insurance company.
- Claims are submitted to hybrid framework where ECM is followed by SVM to detect these fradulent claims.
- There is an expert who detects these forged claims and investigates further with insurance company.
- The legitimate claims are further processed to insurance company and those claim amount are paid to the pstients.

D. Pseudo Code for the Hybrid Approach:

- For each of the incoming claim, apply ECM to form cluster according to the diesease type.
- Apply SVM to each of the cluster those fall into “legitimate” and “fraudulent” classes.

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

Consider an example of duplicate claims of the new disease.

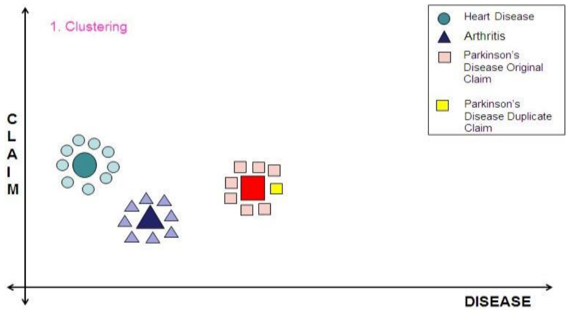
1. Clustering
Heart Disease
Arthritis
Parkinson's Disease Original Claim
Parkinson's Disease Duplicate Claim
CLAIM
DISEASE

Fig.9. Clusters are formed according to disease type.

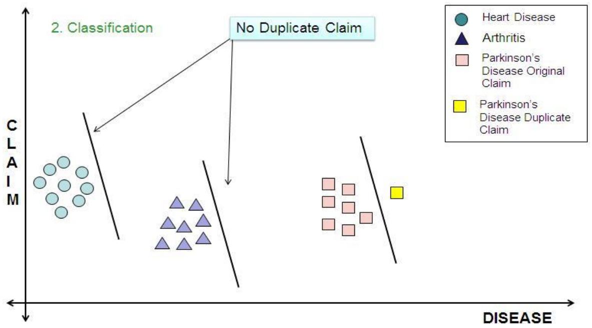
2. Classification
No Duplicate Claim
Heart Disease
Arthritis
Parkinson's Disease Original Claim
Parkinson's Disease Duplicate Claim
CLAIM
DISEASE

Fig.10. Duplicate claim is detected by classyfying it as fradulent 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 further given to SVM algorithm for classification. As result clusterd get formed for all diseases claims including new unknown diseases which wont be possible with traditional clustering method like k-means technique. So, cluster get formed for parkinson’s disease claims as well. Next, the duplicate claims wont get detected on spplying clustering. This drawback is overcome by applying classification based on already formed clusters. Hence, SVM classifies the duplicate claims. Thus this hybrid method using both SVM and ECM shall prove to be more usefull than other methods in medical health insurance domain for detecting health insurance frauds.

V. CONCLUSION

As fraud becomes more complex and the volume of data also grows, it becomes more cumbersome to recognize fraud from bulk of data. We can’t eliminate fraud but we suerly can reduce it. Data moning uncovers pattern hidden in data to deliver knowledge. It invovlves mainly classification and clustering techniques. Cosidering advantages and disadvantages of most of classification and clustering techniques. ECM is chosen as the best clustering technique because it clusters continously flown dynamic data and SVM as best classificaton technique reason being it provides scalabiltiy and usability that are needed in good quality data mining system. Also the quality of generazation and ease of training of SVM is far beyond traditional methods such as neutral methods and radial basis function.

REFERENCES

[1] Dr.Biswendhu Bardhan “frauds in health insurance”.

[2] Melih kirilidoga,Cunyet Asuk (21012) A fraud detection approach with data mining in health insurance.

[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 marix factorization”. The 6th international conference on computer science & education (ICCSE 2011) august 2011. Superstar virgo singapore.

[5] Zhongyuan Zhang, Tao Li,Xhris Ding, Xiangsun Zhang, “Binary Matrix Factorization with Applications”, ’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”.

[7] Haesun park. Class lecture topic:” nonnegative matrix factorization for clustering”. Georgia Institute of Technology Atlanta, 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 K-means Algorithm”,Australian journal of Basic and Applied sciences, 7(8): 140-144,2013 ISSN 1991-8178.

[9] Leonard Wafula Wakoli. “APPLICATION OF K-MEANS CLUSTERING ALGORITHM IN MEDICAL CLAIMS FARUD/ABUSE DETECTION”, MSc Thesis, jomo Kenyatta University of Agriculture And Technology, 2012

[10] Qun Song, Nikola Kasabov, “ECM- A novel online, Evolving clustering method and its applications”, Department of Information Sciene, University of Otago.