Detection of fraudulent activities in Health Insurance using Data Mining

*Poonam Temgire, #Shweta Yendhe, $Damayanti Sonkamble, ¥Prof. Mansi Choche

*,#,$,¥BE (Information Technology), K.C. College of Engineering, Mumbai, India.

*poonamtemgire11@gmail.com, #shwetabyendhe1995@gmail.com, $damayanti423@gmail.com

***Abstract*—Health Insurance Frauds are spread widely and causes huge economic losses to the healthcare insurance companies. Such Fraud relate intentional misleading or misrepresentation intended to result in an unauthorized benefit. Although they make up only a small fraction, such fraudulent claims bring a very high price tag. The prevalence of health insurance frauds keeps proliferating year on year. In order to find and avoid such frauds, data mining tactics can be employed. This includes data extracted from some prior knowledge of health care system and its fraudulent behaviors, analysis of the characteristics of health care insurance data. Data mining is branched into two learning tactics viz., supervised and unsupervised, both of which can be used for fraud detection, by combining the advantages of both the tactics, a hybrid approach for detecting fraudulent claims is proposed.**

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

I. Introduction

Frauds do exist wherever it involves financial transactions. A tempting target for such fraudsters is the health care sector. When health services are provided to the insured people, a set of claims are submitted to the insurance company for reimbursements. Health insurance is similar to any other types of insurance services where a system processes and arbitrates the claims filed to determine if a claim should be funded or by how much.

Health insurance fraud is an act of deluding, or misrepresenting information that results in some sort of benefits to the people involved in the fraud. Main objective behind such frauds is financial benefits.. The net effect of such fraudulent claims is excessive billing amounts, high costs per patient, more digits of patients, higher tests per patient, and so on. In the United States of America, total expenditure is a massive $2.9 trillion, or 18% of GDP. It is unsure of how much amount out of that is pilfered. Approximately 15 per cent of total claims are false claims as estimated. As per reports by Insurance companies in USA approximately 1.6 million insurance claims are filed, and incurs losses totaling up to $34.4 billion annually to healthcare insurance frauds. The statistics is alarming in developing country like India as well. Insurance fraud cases in India surged 21% in 2017. The report advocates that the healthcare industry in India is losing almost Rs.800-Rs.900 crores subjected on fraudulent claims annually.

Health insurance sector is suffering huge losses with very high claims ratio. So, there arises a necessity to minimize or elude fake claims to make health insurance industry exempt from such fraudulent claims arriving through health insurance.

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

- Billing for services not used:

Insurance company being billed for things that never took place. Example: Generating bills for services not rendered by forging signatures of those who provide the invoice.

- Upcoding or unbundling of services:

Insurance company being billed for services that are more expensive than the actual procedure that had been performed. Suppose 30-minute laser therapy session being billed as 60-minute session

- Upcoding or unbundling of items:

Insurance company being billed for medical and surgical accessories and equipment that are costlier than the actual equipment. suppose Billing for Ultrasonic nebulizer while giving the patient a cheaper alternative atomizer nebulizer or billing for a wheelchair while providing the patient for a manual one.

- Twin claims:

Claim submitted by the provider for the exact same service which was provided to a particular individual on a specified date of service that was included in a previously submitted claim. Not submitting exactly the same bill, but manipulating some small parts like the date in order to charge insurance company twice for the same service rendered.

- Not required services:

Registration of claims which by no means is applicable to the condition of a patient.

II. Data Mining

The colossal usage of computers has provided an extra ordinary amount of data at one’s disposal. Due to availability if such tremendous amount of data, experts are facing challenges in excerpting meaningful information from it. Data mining is the process in which information which is hidden, not priorly known and is probably useful,

DOI : 10.18231/2454-9150.2019.0109

is extracted from large databases. Data mining can also be understood as finding the correlations in relational databases and analyzing it. Such potential of data mining can help the health insurance companies to identify fraudulent claims and save the companies from incurring huge losses to such frauds.

Data mining automatically filters from huge amounts of data to extract known/unknown patterns, find new perceptions and then make predictions based on such information learned. Data mining tactics follow the two approaches to learn data mining models i.e. supervised learning, unsupervised learning, and they are explained below:

A. Supervised Learning:

Supervised learning describes itself as machine learning task in which we perceive that a activity design an absorption to an productivity based on sample sets of absorption-productivity sets. It analyzes the provided training data and yields a derived function, which could further be used for mapping new samples .In the supervised learning tactic the model is trained making use of pre-defined class labels. In the context of fraud detection in health insurance the class labels can be given to be classified as either ‘legitimate’ or ‘fraudulent’ claims. The training dataset is employed to build the model. Then any new claim will be compared with the already trained model to make a prediction of its class. A claim will be classified as a legitimate claim if its pattern matches with the pattern of a legitimate behavior else it will be classified as fraudulent.

The advantages are that supervised learning gives more accurate results than of unsupervised learning and depends heavily on training of the data. All classes are meaningful and it can be easily used for pattern classification.

The disadvantage is the difficulty arising while grouping class labels. Moreover, when there is high volume of input data, to label all of them is costly, and accurate identification of claims is necessary because false positives and true negatives can make a bad impression about the insurance company in the perception of its customers. Supervised learning models cannot identify new types of frauds which does not hold similarity in patterns to claim previously classified as fraudulent. Moreover deriving the labeled training samples which will be employed to construct the model requires compelling efforts from the experts.

B. Unsupervised Learning:

Unsupervised learning is a machine learning algorithm used to draw interpretations from datasets constituting of input data without labeled responses.

The most common method of unsupervised learning is array analysis, which is used for exploring and analyzing data to find hidden patterns or arrays in data. As unsupervised learning has no class labels, it focuses on identifying those instances which show exceptional behavior. Unsupervised learning tactics can discover both old as well as new types of frauds since they are not constrained to the fraud patterns which already have previously defined class labels like supervised learning tactics do.

The advantages are it uses exploratory analytics and aims detection of any characteristic which does not abide by the normal behavior and because the approach is not guided it can also explore and find patterns that have not been previously recorded..

While the disadvantage being because of lack of direction, there may be times when the set of features selected for the training provides no interesting knowledge that is discovered. Since both tactics have its own set of advantages and disadvantages, by combining the advantages of both the tactics, a novel hybrid approach for detecting fraudulent claims with greater accuracy can be employed.

III. LITERATURE REVIEW

There are various supervised and unsupervised data mining tactics existing out of which the following are listed.

ANOMALY DETECTION:

In data mining, outlier exposure is the identification unusual case or observations which brings uncertainty by alter significantly from the large amount of the data. In case of health insurance, anomaly detection tactic can be used to calculate the probability of each claim to be fraudulent by examining the previous insurance claims.

Srinivasan et al. [2] proposed the anomaly detection tactic by applying Rule-based Data Mining, an unsupervised tactic, on the insurance claims data collected from Medicare data. Applications that makes use of big data for analyzing health insurance claims to detect fraud, abuse, waste, and errors were devised. Health insurance claim anomalies were detected using these applications that enable private health insurers identify hidden cost overruns that transaction processing systems fail to detect.

SUPPORT VECTOR MACHINES:

In data mining, support vector machines are supervised learning models with correlated learning algorithms that analyze data which is further used for alloting and regression analysis. SVM is principally a classification tactic. The system is trained for determination of a decision boundary between classes of “legitimate” and “fraudulent” claims. Each claim is then compared with the decision boundary and is classified into either legitimate or fraudulent class.

Charles et al. proposed the liblinear implementation of linear SVM [3]. Linear SVM is well known for scaling of features and observations. They incorporated this tactic to a automated bill processing system used in conjunction with human auditing to provide a complete cost regulated solution for medical claims.

NON-NEGATIVE MATRIX FACTORIZATION:

Non-negative Matrix Factorization (NMF) is a feature extraction algorithm. NMF is useful when there are many attributes and the attributes are uncertain and have weak predictability. By combining such attributes, NMF has the ability to produce meaningful patterns, and relationships. This tactic can be used in Health Insurance fraud detection by arraying medical treatment items into several arrays according to usage by different patients. Each array can be shown as group of medical/surgical treatment equipments or items for curing symptoms of similar diseases. Now, this tactic can identify fraud if a medical equipment deviates from one array to another in a period of a month. Its disadvantage being it is intractable to solve.

DOI : 10.18231/2454-9150.2019.0109

Shunzhi Zhu[4] et al. proposed the Non-negative Matrix Factorization(NMF) conducted on a practical dataset supplied by Health Insurance Management Center of Xiamen. They put forward a tactic for arraying semantic features in a prescription collection and group the medical treatment items into groups based on shared semantic features with preservation of natural data non-negativity.

k-MEANS ALGORITHM:

k-means is one of the simplest unsupervised learning algorithms that uses vector quantization to solve the arraying problem. The k-means algorithm takes the parameter k (no of arrays to be formed) as input, and divides a set of n objects into k arrays such that that it results into high intra-array similarity while the inter-array similarity is low. This algorithm the digits of arrays is pre-defined. This turns out to be the drawback for arraying new incoming objects since the digits of arrays is fixed.

Stephen[5] et al. proposed the improved K-means arraying algorithm to solve the segmenting problems of the different health insurance claims. They achieved classification accuracy on iris dataset- 89.33% for Real time Assignment Kmeans (RAK) and 88.67% for Modified-Initialize K-means (MIK) and Traditional KM while the classification accuracy achieved on health insurance claims dataset as 68.1403% for RAK, 68.1402% for MIK and 68.1401% for TKM.

OUTLIER DETECTION: Here, a baseline of the unusual behavior of dosage of medicinal usage for patients is recorded. Any deviation from this baseline indicates the presence of an outlier. It generally results from arraying.

Approach	Advantages	Disadvantages
Anomaly Detection	-testing phase is fast. -purely data driven	-does not give meaningful anomaly score. -complexity in testing phase
Support Vector Machine(SVM)	-scales relatively well to high dimensional data. -generalized, risk of overfitting is less.	-training phase is time consuming. -difficulty in choosing the kernel function.
Non-negative matrix factorization(NMZ)	-Results are easier to interpret. -Provide an additive basis to represent the data.	-There's no hierarchy in the basis. -There's no hierarchy in the basis.
k-Means	-K-Means may be computaHonally faster than hierarchical arraying. -k-Means may produce tighter arrays than hierarchical arraying.	-sensitive to scaling.

IV. PROPOSED SYSTEM

Big limitations of supervised and unsupervised tactics are that the supervised methods cannot classify claims of an remote disease while the unsupervised methods can't detect irregularity when duplicate claims i.e. claims with different dates are filed. So, we propose a hybrid model for the detection of health insurance frauds and to flag them for more inspection. For this, we have chosen Evolving arraying Method (ECM) for arraying purpose since the data we are going to use is dynamic and new data is generated frequently. We chose Support Vector Machine (SVM) for allocation purpose. In this approach, firstly, the insurance claims are grouped into arrays according to the disease type and then they are classified to detect any duplicate claims.

Evolving arraying Method (ECM):

ECM is used to arrayize dynamical data. Dynamical data are the data which keep on changing frequently with regard to time. This is based on the concept of dynamic affixing and modification to the arrays as new data is represented. The modification to the arrays affects both the position of the arrays and the array size, in terms of a radius parameter associated with each array that determines the boundaries of that array. Firstly, the array radius is set to zero. The radius of the array gradually increases as more data points are added to that array. One more parameter is the distance threshold, which determines the affixing of arrays. If the threshold value is small then, there will be more digits of smaller arrays and if the threshold value is large, then there will less digits of larger arrays. Threshold value is selected on the basis of the heuristics of the data points.

Algorithm:

```
    Create the first array centre from the first example I 0
    for each subsequent vector I n do
    Find the minimum distance D min between I n and
each array centre C n
    if D min is less than any array radius then
    Add I n to the nearest array
    else

FInd the array a with minimum value of S i , j , where S i ,
j = D i , j + R i , j , D i , j is the distance between the array
centre and vector j , and R i is the radius of array i
    if S i , a > 2 D  then
    Create a new array
    else
    Update a
    end if
    end if
    end for
```

When array a is updated, its centre is shifted closer to I n and its radius R a (t + 1) is set according to the equation below:

R a (t + 1) = S i , a 2

The new mid of a , C a (t + 1) is set so that its span is on the line between C a (t) and I n at a span of R a (t + 1) .

SVM Steps:
1) Practice (reprocessing Step):

- Explain the two class mark i.e. "legitimate" or "fraudulent"
- Codify the claims filed into two classes using the training data set.
- Make choice of base vectors and find the maximal marginal hyper plane that separates the claims into two classes.

2) Classification:

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

DOI : 10.18231/2454-9150.2019.0109

Steps in Hybrid Model Construction:

- Doctor bills patients for the services/equipment provided to them during their treatment.
- Patient files claims to the insurance provider.
- Claims are submitted to the Hybrid Framework wherein arraying is followed by classification to detect the claims if fraudulent.
- Claims classified as fraudulent are flagged for further investigation with the insurance company.
- The legitimate claims are further passed on to the insurance company and those claims are paid for the patients.

V. CONCLUSION

This paper lists and provides comparative analysis of various approaches for detecting fraudulent behavior in health insurance claims. We have analyzing the mentioned tactics, their advantages and disadvantages in our application for health insurance claim fraud detection. Also we have proposed a hybrid approach for health insurance fraud detection by combining the features of both supervised and unsupervised learning for greater efficiency.

REFERENCES

[1] Fraud Detection in Health Insurance using DataMining tactics

[2] Srinivasan, Uma, and Bavani Arunasalam. "Leveraging big data analytics to reduce healthcare costs." IT professional 15, no. 6 (2013): 21-28.

[3] Using Support Vector Machines to Detect Medical Fraud and Abuse Charles Francis, Noah Pepper, Homer Strong USA, August 30 - September 3, 2011

[4] Zhu, S., Wang, Y., & Wu, Y. (2011). Health care fraud detection using nonnegative matrix factorization. 2011 6th International Conference on Computer Science & Education (ICCSE).

[5] Development of improved K-means arraying for health insurance claims 1 Stephen G. Fashoto, 2 Adekunle Adekoya, 3 Jacob A. Gbadeyan and 4 J. S. Sadiku

DOI : 10.18231/2454-9150.2019.0109