A HEURISTIC APPROACH TO PREDICT FALSE CLAIMS IN HEALTH CARE SECTORS

[1] Survase Komal, [2] Sangshetty Manisha, [3] Pawar Smita , [4] Waghire Rajani, [5] Prof.P.B.Jawalkar
[1]UG student, [2]UG student, [3]UG student, [4]UG student, [5] UG student Guide,
[1]Computer Engineering,
[1]JSPM'S BSIOTR, Pune, India

Abstract— Due to happening of health care fraud there is huge loss for companies. To overcome this problem some systems need to implement to identify fraud. To detect health care fraud many systems uses synthesized datasets, data mining techniques and hybrid approach. There are many health care fraud detection systems are available i) Survey on Hybrid Approach for Fraud Detection in Health Insurance, ii) Data Mining for Fraud Detection, iii) A survey on statistical methods for health care fraud detection. Many systems related to health care fraud are having performance issues regarding detection of fraud, so this paper proposes an idea of health care fraud detection. To enhance the process of fraud claims detection of the doctors at the insurance company's end proposed method put forwards an idea of identifying fraud claims by clustering the claims based on the protocols by using the C-means clustering technique which is then powered with Hidden markov model to extract the fraud list and this process is catalyzed by fuzzy logic classification theory.

Keywords— Protocol Collection,C Means Clustering, Hidden Markov Model, Fuzzy Classification.

I. INTRODUCTION

Frauds exist wherever when it involves money transactions. in health care fraud detection system , identifies the fraud claims of doctor at insurance company end. there are many insurance companies are existed to provides health claims facility to the patients but these companies suffers from different fraud claims by doctors. for example, In the United States, total health spending in America is a massive $2.7 trillion, or 17% of GDP. No one knows for sure how much of that is embezzled, but in 2012 Donald Berwick, a former head of the Centres for Medicare and Medicaid Services(CMS), and Andrew Hackbarth of the RAND Corporation,estimated that fraud (and the extra rules and inspections required to fight it) added as much as $98 billion, or roughly10%, to annual Medicare and Medicaid spending, and up to $272 billion across the entire health system [1].

This paper will discuss the claims data that are processed and detect the fraud claims by the doctors. Using post-payment claims data, we can perform many types of data analytics and data mining techniques to identify potential frauds. There are many types of insurance frauds, the following is a list of frauds in health insurance that are most commonly mentioned:

1. Billing for services not rendered.
2. Billing for a non-covered service as a covered service.
3. Number of Recalls.
4. Recalls in period.
5. Misrepresenting providers of service.
6. One doctor reffer to another doctor.
7. Incorrect reporting of diagnoses or procedures.
8. suggesting dietory

In this paper, we developed algorithms that target at one type of frauds. That is the suspicious provider communities that either share patients between or refer patients to each other. there are two techniques are such as fuzzy c mean clustering and hidden morkov model. Firstly , in fuzzy clustering, each point has a probability of belonging to each cluster. This algorithm belongs to the family of fuzzy logic based clustering algorithms. Here all the data that is been collected for the calming of insurance is clustered logically using c means clustering. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula.

II. LITERATURE SURVEY

[1]Surveys different data mining techniques which is applied to the application. health insurance fraud is an intentional act which make financial benefit to individual or group. to detect all the fraud claim it uses data mining techniques which is divided,such as supervised and unsupervised learning techniques.

[2]Explains supervised and unsupervised algorithms to detect the fraud at insurance company end. the major drawback of both these algorithm is, they cannot classify or divide the fraud claims as per diseases.Evolving Clustering Method (ECM):

ECM is used to cluster or divide the dynamic data. dynamic data means ,the data which keep changed with respect to time. when new data set comes in system, ECM clusters that data by modifying the size and position of cluster.

support vector machine(SVM):

The support vector machine is a supervised learning technique used for classification. It has an initial training phase.

[3]Focuses on frauds and uses techniques such as Data mining empowers a variety of insurance providers with the ability to predict which claims are fraudulent so they can effectively target their resources and recoup significant amounts of money.

[4]Narrates insurance companies use human inspections and heuristic rules to detect fraud. First, it is impossible to detect all emerge constantly. health care fraud by manual inspection over large databases.

DATA UNDERSTANDING AND PREPROCESSING:

The storage and processing of this data was conducted within the insurance company's infrastructure in compliance with privacy regulations.

EMPIRICAL CLUSTERING RESULTS:

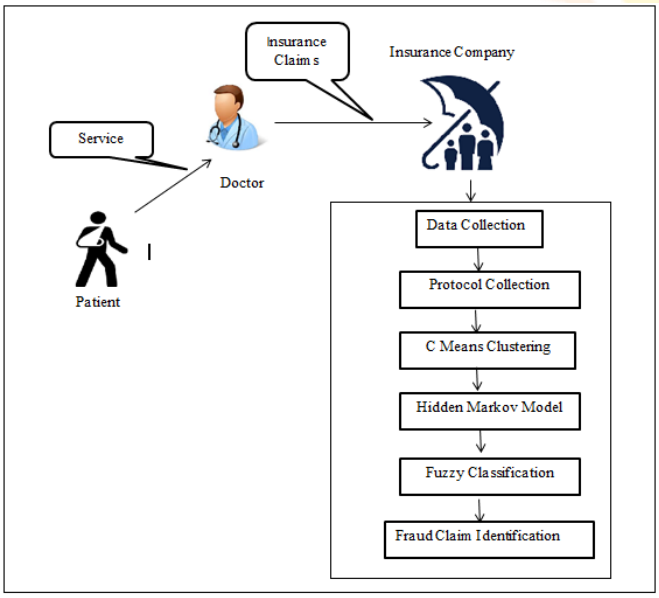
Most clustering algorithms require the users to input the number of clusters they desires. Therefore, we conduct a series of clustering experiments and discuss the results with business experts to decide which numbers of clusters
are desirable.

[5]Introduce an effective medical claim fraud/abuse detection system based on data mining used by a Chilean private health insurance company. There is a difference between fraud prevention and detection Data mining which is part of an iterative process called knowledge discovery in databases (KDD) can assist to extract this knowledge automatically.

The performance of a k-Nearest Neighbor (KNN) algorithm with the distance metric being optimized using a genetic algorithm was applied in a real world fraud detection problems faced by the HIC. and fuzzy logic and genetic algorithms are used. The following few methods are used [A] Entities and Medical Claim Data,[B]. Business and Data Understanding,[C]Data Preparation, [D]Modeling [E]Incorporation into Fraud Detection Workflow.

[6]In this paper, we introduce different methods and techniques to detect it. There' use multilayer perceptron neural network, Neural Nets (NN), Bayesian Nets (BN), Naive Bayes (NB), Artificial Immune Systems (AIS), Decision Trees (DT), this are the Different Techniques for Fraud Detection. The aims of this paper are to assess the use of technique of decision trees. In combination with the management model CRISP-DM, this technique helps in the prevention of bank fraud .

III. PROPOSED SYSTEM



Proposed Framework is been presented in above figure 1: and is been detailed as below sequentially.

Phase 1: Data gathering

Two Dimensional vector is ben generated consisting of all records from data base for future evaluation.

Phase 2: Protocol writing

Correct set of Rules have been stored in rule base based on expert knowledge for better processing. A multidimensional vector has been used to store all rules.

Phase 3: C Means Clustering

Here every one of the information that is been gathered for the quieting of protection is bunched legitimately utilizing c implies grouping with the accompanying procedure. This calculation works by doling out participation to every information direct relating toward each bunch focus on the premise of separation between the group focus and the information point. Progressively the information is close to the bunch focus more is its enrollment towards the specific group focus. Unmistakably, summation of enrollment of every information indicate ought to be equivalent one. After every emphasis enrollment and group focuses are refreshed by procedure

Phase 4: HMM(hidden Markov Model)

Markov Model(HMM) is an intense measurable device for displaying generative successions that can be described by a fundamental procedure creating a discernible arrangement. Gee have discovered application in numerous ranges inspired by flag handling, and specifically discourse preparing, yet have likewise been connected with accomplishment to low level NLP undertakings, for example, grammatical feature labeling, express lumping, and extricating target data from archives. Andrei Markov gave his name to the scientific hypothesis of Markov procedures in the mid twentieth century however it was Baum and his partners that built up the hypothesis of Well in the 1960s.

Phase 5: Fuzzy Classification

Here in this step based on results from forward probability and backward probability evaluation Transaction vectors are been generated . Fuzzy Classification is been applied to given claims and transaction vector is been used fro comparison. Data is been distributed based on range and classified as low to very high fraud.

Phase 6: Fraud Claims Identification

All Claims are been displayed based on level of fraud identified for them and then marked as processed

---

Algorithmic 1: steps for Fuzzy c-means clustering

Let X = $\{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and V = $\{v_1, v_2, v_3 ..., v_c\}$ be the set of centers.

Step 1: Randomly select *'c'* cluster centers.

Step 2: Calculate the fuzzy membership $'\mu_{ij}'$ using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

Step 3: Compute the fuzzy centers $'v_i'$ using:

Repeat step 2) and 3) until the minimum 'J' value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.

Where,

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

'k' is the iteration step. *'β'* is the termination criterion between [0, 1]. *'U = $(\mu_{ij})_{n*c}$'* is the fuzzy membership matrix. *'J'* is the objective function.

Algorithm 2: the Algorithm to Calculate A Count and Percentage Matrix

---

Required: Two Inputs

a. A health care claims dataset. It is a transaction database with each record is a transaction or a claim.

b. Size of Each Batch (Size), this gives the size of each batch that needs to be evaluated.

1. Separate all providers into a list of batches according to Size and read them into a macro variable, Batch List
2. for all Batches (B) in Batch List do
3. for all providers in B do
4. calculated pairwise count between providers in each batch
5. end for
6. Save the batch results
7. end for
8. Aggregate batch results into one table (Count Matrix)

9. for all Providers in variable List do
10. calculated percentages of each pair of providers
11. Save the batch results into one table (Percentage Matrix)
12. end for
13. return A Count Matrix and A Percentage Matrix

Algorithm 3: HMM Pseudo Code

Step 0 : A first-order Markov chain generates the hidden state sequence (path): initial state probs:
Step 1: A set of emission / output distributions A j (•) (one per state) converts this state path into a sequence of observations yt.
Step 2: Even though hidden state sequence is first-order Markov, the output process may not be Markov of any order
Step 3: Discrete state, discrete output models can approximate any continuous dynamics and Observation mapping even if nonlinear; however this is usually not practical

IV. RESULTS AND DICCUSSION

Proposed system of traffic symbol detection system detection is deployed as a standalone system using Apache Tomcat.

Performance is evaluated based on the precision and recall parameters. Precision is defined as the ratio of number of relevant traffic symbols are detected to the total number of relevant and irrelevant relevant Health frauds are detected. Relative effectiveness of the system is well expressed by using precision parameters.

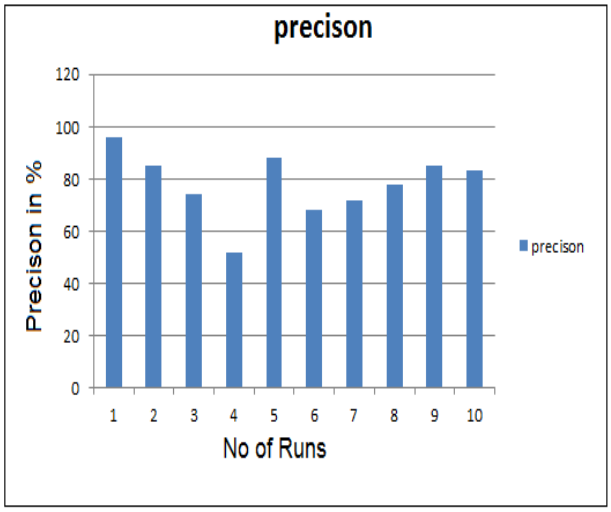
Whereas the recall can be defined as the ratio of number of relevant frauds are detected to the total number of relevant frauds are detected not detected. Absolute accuracy of the system is well narrated by using recall parameters.

System can be evaluated using precision and recall parameters, and they can be more clearly elaborated as follows.

- X = The numbers of relevant health fraud are detected,
- Y = the number of relevant health frauds are not detected, and
- Z = the number of irrelevant frauds are detected.

So, Precision = (X/ (X+ Z))*100
And Recall = (X/ (X+ Y))*100

Fig.2. Average precision for Health care fraud



In Fig. 2, by observing it is clear that the average precision obtained for Health fraud detection correlation mechanism is approximately 84%.

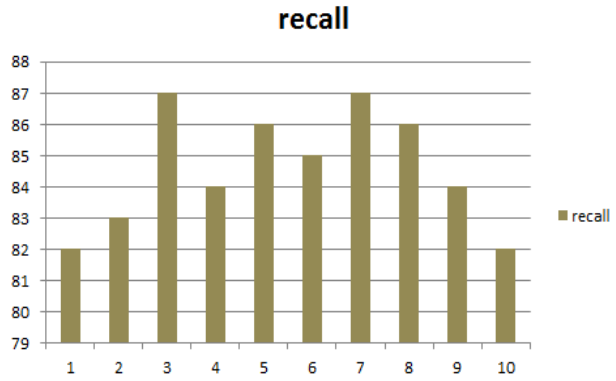


Fig.3. Average Recall for health care fraud system

Figure 3 shows that the system gives 84.66% recall for the Health care fraud technique using image correlation mechanism. By comparing these two graphs we can conclude that the health care fraud .

V. CONCLUSION AND FUTURE SCOPE

Successful Protocol Definition for Fraud Scenarios. Accurate Fraud Claims by the doctors Whole scenario will be implement in proper web paradigm. Future System can be implemented for more number of protocols . The implemented system could be released as API and could be helpful to every insurance company.

REFERENCES

[1] "Survey on Hybrid Approach for Fraud Detection in Health Insurance",Punam Devidas Bagul, Sachin Bojewar and Ankit Sanghavi,[2016].
[2] "Fraud Detection in Health Insurance using Data Mining Techniques",Vipula Rawte and G Anuradha, [2015].
[3] "Data Mining for Fraud Detection",Manjunath K.V and Patharaju S.D, [2015].
[4] "Application of Clustering Methods to Health Insurance Fraud Detection",Yi Peng,Gang Kou, Alan Sabatka2,Zhengxin Chen,Deepak Khazanchil and Yong Shi
[5] "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile",Pedro A. Ortega, Cristi´an J. Figueroa and Gonzalo A. Ruz
[6] "Survey on Fraud Detection Techniques Using Data Mining",Muhammad Arif andAmil Roohani Dar.
[7] "Using Data Mining to Detect Health Care Fraud and Abuse", Hossein Joudaki, Arash Rashidian, Behrouz Minaei Bidgoli, Mahmood Mahmoodi, Bijan Geraili, MahdiNasiri,[2015].
[8] "DEA restricts narcotic pain drug prescriptions", Radnofsky, L., and Walker, J., 2014.