A Model System to Identify Health Care Frauds

N. I. Ujloomwale[1] Aditi Kamath[2] Shraddha Hundalekar[3] Pooja Divase[4] Darshana Akadkar[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Modern Education Society's College of Engineering, Pune, India

Abstract— As the human life marches towards the modern amenities, all the sectors of life become more and more advanced, Health care is not spared from this. The revolutionary health care policy concept eventually facilitates all the patients irrespective of any cast and creed to avail the best services of the doctors for their diseases. Many of the health care insurance companies are existed to provide this facility for the peoples, but all of them are suffering from the headache of fraud insurance claims from the doctors. Many systems are existed to deal with these kinds of fraud health insurance claims from the doctors, but most of them are not up to the mark to identify the proper fraud detection operandi. So as a small step towards this, the proposed system develop a web application panel for both the doctors and insurance companies to identify the fraud claims of the doctors at the insurance company's end using Artificial Neural Network which is powered with fuzzy classification.

Key words: Clustering Algorithms, Matrix Converters, Detection Algorithms, Insurance, Algorithm Design and Analysis, Drugs (Medicines), ANN, K-Means Clustering, Fuzzy Logic

I. INTRODUCTION

Purposefully deceiving the health care insurance company results in the benefits being misused by certain group of people is described as a Health Insurance Fraud. These frauds are brought into action in-order to take control over the benefits deployed to the common people.

In the recent years there has been a tremendous increase in the medical fraud all over the globe. According to various surveys and reports the health insurance industry in India is losing approximately Rs. 800 crores on fraud claims on fraud claims per year. Health insurance is a large sector with a very high claim ratio. In-order to safeguard the health insurance industry of the increasing level of frauds, it is very necessary to eliminate the fake claims available at the insurance companies. Also, there can be seen a sudden rise in the different technologies,

Data mining techniques along with Data Analysis is an upcoming technology, which is becoming more popular in the health care insurance industry, as it effectively helps in detecting the unknown and more valuable information in a more efficient and accurate way.

Hence the document A Model System to Identify Health Care Frauds describes the issues related to the system and what actions are to be performed in order to come up with a better solution.

II. RELATED WORK

The need to develop an efficient fraud detection model is a necessity in today's era with the increase in the population and the infrastructure. The different Data mining concepts like K-means Clustering, Community algorithm, Artificial Neural Network as well as fuzzy logic was brought into use to develop different systems which could accurately and efficiently identify the differences and complex fraud claims. A brief survey on all these approaches have been discussed below:

The earliest existing system proposed in [2] explains an idea of Health detection on the basis of the community based algorithm. This community based algorithm mainly detects the relation between physicians. This paper focuses on the patients shared between physicians or the number of patients treated by more than two doctors. It uses two types of algorithm, first is to calculate a count percentage matrix in which a matrix is created which consists of the count of patient and doctor. It creates a count ratio based on how many time patient visits a doctor. In the second algorithm, it determines the relationship between doctors. It finds community between them. It is mainly used to detect the reference between the doctors and finds which doctors are suspicious to be in a community. The proposed method has a limitation as it doesn't check for accuracy of the results. It focuses on 100% test results. It considers only two types of relationships between physicians.

The proposed work in [3] introduces an idea of Health Care Fraud Detection as supervised and unsupervised algorithms. It consists of Evolving Clustering method (ECM), Support Vector machine (SVM), Anomaly Detection and K-means algorithm. With the help of anomaly detection technique, probability of fraud claims is calculated by analysis of previous claims and records. Using ECM dynamic data (continuously changing data with respect to time) clustering is done. It makes use of a parameter named as a distance threshold (DTHR), which is used to determine the sum of clusters. The lesser number of threshold values, more the number of small clusters and larger threshold value results as less number of large clusters. SVM is a classification technique which has an initial training phase where already classified data is fed to the algorithm. In this approach, firstly the insurance claims are clustered on the basis of disease type and then classified to trace frauds. But the problem with this model is, when large datasets are to be processed, then this algorithm doesn't work properly. Also, if there is any unknown disease, then it cannot detect fraud there.

The author in [4] narrates an idea to detect Frauds in the health care system on the basis of Nonnegative Matrix Factorization algorithm (NMF). This algorithm is a clustering approach to understand the latent structure of the observed matrix. NMF respects the non-negativity that is inherent in all the health care data sets. For clustering it makes use of NxM matrix, where N represents the number of patients received medical therapies in one month and M stands for a number of medical treatments listed in all prescriptions in one month. The measurement vectors, i.e vt is arranged into the columns of the NxT matrix. The relationship between V, W and H is written as V^WH. At the end of every month, shift track of each patient is detected on the basis of clustering results. Patients who induces the medical treatment items shift frequently from one cluster to another, is considered as a

fraud suspicious patient. But to use NMF in a large dataset, a matrix with millions of rows and columns is needed which is practically not possible. Hence it is only applicable for small data sets.

The system described in [5] expresses a Novel approach to Health care fraud detection. It is using community algorithm based on spectral analysis. It makes use of gap cut algorithm, which depends on spectral analysis. The Gap cut algorithm divides network into an unknown number of communities. In gap cut algorithm, we determine the minimum gap between values in order to separate different communities. Spectral analysis is a result of studies in linear algebra and solutions of systems of linear equations and their generalizations, it extends the eigenvector and eigenvalue theory of a single square matrix to a much broader theory of the structure of operations in a variety of mathematical spaces. The problem with this model is, it needs pre-labelled data and suffer the drawbacks of mis-classifications. But this paper does not focus on the multi - mode network of health care data sets.

The methodology in [6] introduces an idea of Health care fraud detection using C means Fuzzy classification. This paper analyses the types of fuzzy c-means clustering i.e. approximation fuzzy, c-means clustering and lateral c-means clustering on the basis of numerical methods. The literal coding fuzzy, c-means algorithm is compared with a table-driven approach (AFCM). The results of the comparison show that the AFCM takes one sixth less computer time and also provides the same accuracy as the literal implementation. First general c-means clustering algorithm is explained using the mathematical equation to calculate the Euclidean centre of the clusters. In the approximation method, the internal and external tables are drawn using fuzzy set rules. The calculation part is relatively less in comparison to the traditional method, the exponents are used in the starting steps and external tables are drawn and by using appropriate fuzzy c mean clustering reduces the time complexities by approximately 16.25 percent. Both the algorithm gives better results when implemented practically.

III. PROPOSED MODEL

The description of the System Architecture is presented as:

The doctors' claims collected are being pre-processed. In the Pre-processing the removal of noisy data, data integration, normalization, discretization of the data is being done. The pre-processed data are then labelled for the classification. Integer values are used for the data labelling. If the label is 1 then the claim is fraud and if the label is 0 then no fraud. The system should identify the fraud claims and also notify to the users accurately. Data is then clustered into by using K-means clustering. Artificial Neural Network approach is used for passing information in the reverse direction and adjusting the network to react that information. Fuzzy classification uses the fuzzy logic, which is a form of many-valued logic in which truth values of variables may be any real number between 0 and 1. Thus the fraud claim identification is made.

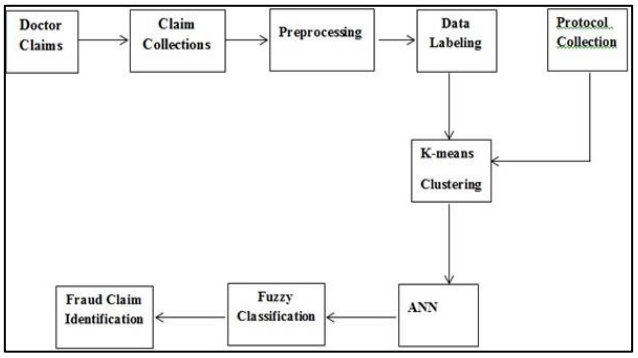


Fig. 1: System Architecture for the Fraud Detection

IV. METHODOLOGIES IN THE SYSTEM EXECUTION

The proposed methodology of the system can be explained through following modules:

A. *Claim Collection*

- Input: Collection claims from doctor
- Process: Preprocess the Claim ID
- Output: Store in a list

B. *Protocol Collection*

- Input: Protocol Collection from Medical Council
- Process: Preprocess protocols
- Output: Store in a list

C. *K Means Clustering*

- Input: Claim list and protocol list
- Process: K Means Clustering
- Output: Clusters

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

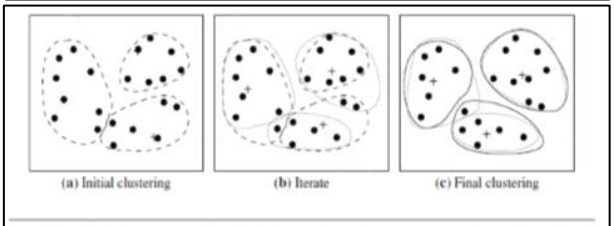
Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;



Clustering of a set of objects using the *k*-means method; for (b) update cluster centers and reassign objects accordingly (the mean of each cluster is marked by a +).

Fig. 2: K-Means Algorithm

D. Artificial Neural Networks

- Input: Clusters
- Process: Neuron Estimations
- Output: Fraud Probability Estimation

Artificial neural networks (ANNs) or connectionist systems are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve performance) to do tasks by considering examples, generally without task-specific programming. ANN is the most powerful learning model. It can have a wide range of complex functions which represents a multi-dimensional input-output maps. ANN is also an information processing paradigm that is motivated by way biological nervous system.

E. Fuzzy Classification

- Input: Fraud probability estimation index
- Process: Fuzzy Crisp values and IF THEN Rules
- Output: Fraud case identification

Rule-based system for classification have certain disadvantages hence by discretizing the values and then applying fuzzy logic to allow "fuzzy" thresholds or boundaries to be defined for different categories. Rather than having a precise cutoff between categories, fuzzy logic uses truth values between 0.0 and 1.0 to represent the degree of membership that a certain value has in a given category. Fuzzy logic is a form of many valued logics in which the truth values of variables may be any real number between 0 and 1. It is employed to handle the concept of partial truth, where the truth value may range between completely true and completely false. Fuzzy logic is closely attached to human thinking and by the way humans express their thoughts.

V. Future Scope

In the near future the proposed system can be enhanced in order to work as an independent Application Program Interface (API). Also, it can be ensured that the system can be equipped to perform adequately in all sectors around the globe. The system can also be enhanced to perform on a number of different protocols.

VI. Conclusion

Health care and insurances are one of the most important aspects of today's life. As the volume of data (insurance claims) increases, the fraud becomes more sophisticated. The fraud, then, cannot be identified from a great bulk of data. But with the help of different Data Mining concepts like K-means Clustering, Artificial Neural Networks (ANN), and Fuzzy Logic Classification the rate of the fraud claims can be minimized to the greatest extent. With the reduced rate of fraud claims, the health insurance companies and the common people can avail the benefits of the insurance claims. Also the country's financial and health care industry is safeguarded from these fraud claims. Thus the idea of developing a model to identify the medical health frauds is beneficial to all as it saves time, money and energy.

References

[1] Hao Peng, Mengzhuo You - School of Computer Science and Technology, National Engineering Laboratory for eCommerce, Shandong University Jinan, China, "The Health Care Fraud Detection Using the Pharmacopoeia Spectrum Tree and Neural Network Analytic Contribution Hierarchy Process",in IEEE TrustCom/BigDataSE/ISPA 2016

[2] Song Chen and Aryya Gangopadhyay, "Health Care Fraud Detection with Community Detection Algorithm", in IEEE 2016.

[3] Vipula Rawte and G Anuradha, "Fraud Detection in Health Insurance using Data Mining Techniques", in IEEE 2015.

[4] Shunzhi Zhu and Yan Wang, Yun Wu, "Health Care Fraud Detection Using Nonnegative Matrix Factorization", in IEEE 2011.

[5] Song Chen and Aryya Gangopadhyay, "A Novel Approach to Uncover Health Care Frauds through Spectral Analysis", in IEEE 2013.

[6] Robert L. Cannon, Jitendra V. Dave, and James C. Bezdek, "Efficient Implementation of the Fuzzy c-Means Clusteng Algorithms", in IEEE 1986.

[7] Shweta Taneja, "A new approach for data classification using fuzzy logic", in IEEE 2016.

[8] Md. Ra_ul Hassan and Baikunth Nath, "Stock market Forecasting Using Hidden Markov Model: A New Approach", in proceedings of the international conference on ISDA, 2005.

[9] S. Patil and V. Bhusari, "Study of Hidden Markov Model in Credit Card Fraudulent Detection", in IEEE 2016.

[10] Mrs. Dharani. S, Mrs. Shoba. S.A. - Assistant professor, Department of Computer Science Arcot Sri Mahalakshmi women's College, Villapakkam, Tamil Nadu, India, "Identifying The Fraud Detection In Health Care System Using Data Mining", in International Research Journal of Engineering and Technology (IRJET).

[11] Jing Li & Kuei-Ying Huang & Jionghua Jin & Jianjun Shi - Department of Industrial Engineering, Arizona State University, "A survey on statistical methods for health care Fraud detection" in Springer Science Health Care Manage Sci 29 May 2007.

[12] Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri2 & Mohammad Arab - Department of Health Management and Economics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran, "Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature", Vol. 7, No. 1, in Global Journal of Health Science 2015.

[13] Lutfun Nahar Lata Department of Computer Science and Engineering Ahsanullah University of Science and Technology Dhaka-1208, Bangladesh, "A Comprehensive Survey of Fraud Detection Techniques", in International Journal of Applied Information Systems (IJAIS).

[14] Emanuel Mineda Carneiro, Luiz Alberto Vieira Dias; Adilson Marques da Cunha -Departamento de Ciencia da Computação Instituto Tecnologico de Aeronautica

(ITA), "Cluster Analysis and Artificial Neural Networks - A Case Study in Credit Card Fraud Detection", in 2015.

