A Model Approch to Discover the Health Care Fraud Claims

Mayuresh R. Bodake[1] Keval B. Randive[2] Prof. A. S. Hambarde[3]
[1,2,3]Department of Computer Engineering
[1,2,3]KJ College of Engineering and Management Research, Pune Savitribai Phule Pune University, India

Abstract— In many advanced countries like USA, Europe and many more, including India is providing health insurance facility for their citizens. In this facility a patient need not to give any kind of upfront fees to the Doctor to get the desired services. After having the services doctor will claim his fees with the Insurance Company with all the details of the services provided to the patient. Here some Doctor may fall in greed and they may charge excess than that of actual charges, which is actually a fraud which gives rise to unethical practices. Some methodologies are existed to estimate the fraud claims raised by the Doctors at the health insurance service provider's end. But it is always the question of a precision is raised. So as a tiny step towards this to improve the accuracy this paper proposes an idea of using the concept of Machine learning like K- Nearest Neighbor and Artificial Intelligence. This process is powered with Entropy Analysis and Logistic Regression.

Key words: K-Nearest Neighbor, Entropy Estimation, ANN, Correlation Estimation

I. INTRODUCTION

KNN stands for K nearest neighbours it is an important technique in the realm of Machine Learning. K nearest neighbour is a classification technique and is widely used for performing machine learning tasks. It should not be confused with the K-means algorithm which is utilised in the big data field for clustering. K-nearest neighbour is a supervised algorithm which classifies the data elements within the dataset by classification of its k-nearest neighbours.

The K-nearest neighbour algorithm classifies the data elements according to labels, these labels are generated and not given explicitly. Therefore, the labels are guessed by the algorithm by other data elements around the target data element and their labels. These neighbouring data elements with their labels are then asked to vote for a label for the target data element. Hence, the neighbours help the algorithm in the labelling of the target element. The K here refers to the number of neighbours that have been referred to vote for the naming of the data element.

K-nearest neighbour is not very complicated to use and can be really useful for generating very accurate results. Due to its simplicity, it has been used extensively in various applications. K nearest neighbour is primarily used for classification problems, but it can also be used for applications in the field of regression. Unlike classification, where the values were voted across the k number of neighbours, in regression the average value is selected from the collection of k number of neighbours.

K- nearest neighbour is widely used as an application in the field of machine learning as it does not require any parameters for its execution. It is also not counted among the eager algorithms, that is the K nearest neighbour doesn't not require a large training dataset. This is beneficial as it has a very short and brief training period which is one of the reasons it is one of the most widely used algorithm as it is extremely easy to deploy, and can be done on a short notice. The K- nearest neighbour can also be utilised for applications concerning classification. It is one of the most versatile of the machine learning algorithms. The classification is achieved by a small modification of the algorithm, as it performs the classification by voting done by the neighbouring elements.

Artificial Neural Networks are one of the most essential tool of machine learning. It is widely used in various applications, as it is quite a powerful data modelling tool. The Artificial Neural network is used to find patterns and complex relationships between the values in a dataset, to he outputs or inputs. This is particularly useful as it can process large amounts of information in little time and extract all the common features in the dataset, pertaining to the use case of the algorithm.

The Artificial Neural Network is aptly named, as the technique has been inspired entirely by a human brain. It is modelled after the brain and shares most of its functions too. The most basic unit of the brain is a tiny cell called as a neuron. This neuron is capable of receiving various information through the other neurons and respond if it was activated. The activation only occurs when the conditions and other factors have surpassed the threshold values of the neuron and have fulfilled the conditions for the activation.

The neurons when activated emit a short electrical pulse that travels down the neuron and its axons into the other neurons. The brain consists of billions of these tiny neurons, each one of them acting like tiny switches that are sensitive to the inputs, and when the threshold id breached, they activate by sending an electrical impulse. This process has been emulated in artificial neural networks, by designing a neuron capable of being activated when certain pre-defined parameters that reach a threshold value.

The Artificial neural network if is setup properly, emulates human-like behaviour. In the sense that it can intuitively understand the situation and actually learn like a human brain. This makes it a very powerful and efficient alternative for a random approximation tool. The artificial neural network enhances already available tools for data analysis. Usually the artificial neural networks are layered on top of each other, in a way that the output of one is the input of the other immediate layer. This provides a greater degree of precision and increases the overall performance of the system as a whole.

Artificial neural networks are commonly used in image processing applications, such as analysing medical images, and they are also quite useful for predicting weather conditions over a certain area given the changing variables that can be utilised by the network to provide valuable insight for the weather prediction.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

II. LITERATURE SURVEY

M.Frahadi [1] In the recent years EHR i.e. Electronic Health Record is growing day by day in many countries. There is an improvement in the health sector in quality of health care and clinical visits. The measures and the rules of the EHR is decided by the Health Insurance Portability and Accountability Act (HIPAA) to ensure the integrity, security availability. Due to EHR the data collected includes the information of patients that will help to make the clinical decision in future. The main purpose of this paper is to provide security to EHR application to map the requirements HIPAA.

M. Ahmed [2] The EHR is coming with new mechanisms and components such as investigation of fraudulent medical claims, prevention and detection. Medical theft identification has become one booming topic of the recent years. Large amount of theft was found in America. So EHR should take action regarding this fraudulent activity such as health exchange policies, standards must be protectively prevented. This paper prevents the various attacks which are going to take place in health care sections. It also prevents the malicious attacks which would lead to financial losses.

R.Bauder [3] Healthcare sector is one of the important sectors in peoples' lives. As seen this growth there should be good and proper health facilities provided to patients. This growth of health care comes in financial as it is usually managed by the insurance company. When it comes to financial there are the chances of fraud and the many other activities. Thus, to improve the detection of fraudulent activities the supervised method is implemented which is better than the unsupervised learning.

K. Ng [4] In this paper the data mining technique is used to find the fraudulent activities which are done in the health care sector. Thus, the use of the data mining with his multitude of challenges are presented. The huff model is used to co- operate knowledge with poor data. Before this there were lots of complaints towards consumer frauds and non-compliances came in different forms. This was one of the major complaints regarding the prescription shopping. As the commodities were targeted directly to the doctor and the pharmacies.

P.Bharathi [5] Content Based Image Retrieval Systems (CBIR) is became one of the recent growing sectors in image processing and became one of the recent growing application in the field of medical images. CBIR is also called as query image content.Content Based means when we observe images, every image contains the content such as tags, keyword, and descriptions associated with the images. Thus, in this paper the proposed method is compared with histogram interaction base classification with the KNN classification separately. The proposed method is better thanthe existing techniques.

J. AIKhateeb [6] In this proposed system HWR i.e. Handwriting recognition plays the key role in checking, verification, mail sorting and the various activities of human computer interactions. HWR is divided in two sectors, viz. offline system and online system. The online system is based on the pen movements which depends on dynamic writing and the offline recognition depends on written text image. The words are divided into blocks and then the mean value is computed of the blocks. Then by using KNN classifier the blocks are classified. This method shows the best result, better than the other methodimplemented in past.

I. Fridge [7] Speech is the expression or thought which is conveyed for the interaction purposes. The target of Speech recognition is to decode the speech signal and identify the text pronounced by the speaker. In this paper they have used KNN algorithm for its simple characterization, accuracy,efficient Time complexity and also for the simple execution. Thus, for improvement they have used fuzzy KNN i.e. FkNN to explore the contribution of fuzzy. Thus, the FkNN algorithm as higher capability than the other method of recognition, which confirms the advantage of the fuzzy concept.

T.Dong [8] There has been rapid growth in data mining and network technologies in the recent years. This paper shows that there has been growth in text categorization which is done with help of the KN classification. This plays the key role in government decisions and in the business sector. There have been some shortcomings in traditional KNN. KNN algorithm used in this paper is an improved version of KNN text algorithm.

X. Wang [9] Intelligent Information techniques is one of the most advanced technologies for information processing in recent years.In this paper KNN is used which is based on the machine diagnosing model. Artificial Intelligence is one of the techniques which is used incontext. Most of the pattern classification and recognition tasks are problem oriented. There different method such as hidden Markov model, Bayesian classifier, neural network and also KNN method. Thus the KNN method as got the highest vote for the accuracy of the output.

Ali Muhtar [10] Artificial Neural Network is one of the most used technique in the recent years. ANN is for analyzing, learning, modeling, which is the one of the most complicated phenomenon. In this paper their comparison between ANN using PSO and particle swarm Optimization to propagate the maximum power point tracking in photovoltaic systems. The result showed in this paper by ANN using the PSO as the training algorithm required 17 epochs to converge.

R. Kamesh [11] Model predictive control is used in the industrial applications, such as blinding, mils, boilers, etc. MPC has an ability to provide the solution to feed forward and feedback on the multiple processes on the interactive loops. Thus the MPC is used in industrial application for the most of the time. In this paper the MPC is integrated with the extended Kalman filter (EKF) and it fully depends on data driven artificial neural network. The proposed ANN-EKFMPC is integrated with the proportional integral controller. Due to this approach, there is control in effort minimization.The proposed technology is founded as very versatile as it is dependent purely on the data driven model.

F. Azevedo [12] Artificial Neural Network is later extended to Multilayer perceptron. Multilayer perceptron consists of three layers there are input layer, hidden layer and output layer. The MLP is technique comes under the supervised learning. In this paper cross validation and ROC is used together with standard propagation ANN MLP training algorithm. The main motivation of this paper is to

give quantitative evaluation and a generalization of the knowledge during supervised learning.

M. Liu [13] This paper provides the information about the design and operation of an algorithm to control or monitor the performance of research octane number (RON) testing a double ANN–multidimensional GC. The double ANN regression model was developed between the output of hydrocarbon analysis and determined research octane number (RON). Earlier Partial least square method was developed but it was not successful as a double ANN regression model. The result of double ANN regression model better than PLS.

III. PROPOSED METHODOLOGY

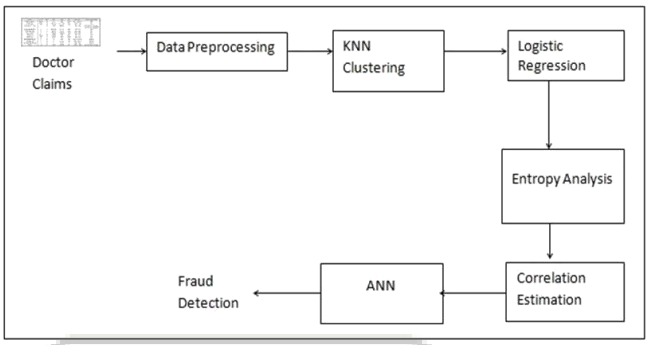

Fig. 1: Overview of Health care Fraud Detection System

The proposed system for detecting Fraud in the health care system by the Doctors can broadly describe as below.

1) Step 1: Data Collection - The proposed model is designed for the two entities like health Insurance company Staff - Receives the claims of many doctors that need to be processed for that instance.

And another one is Health council - Here this entity is defined the rules for the particular disease like suggesting medicine, Number of recalls, Period in between each recalls, Referring Doctor, Referring Doctor Specialty, service cost, suggesting procedures and suggesting dietary.

As the input to the system Health Insurance staff receives the claims need to be processed for that instance in a Workbook format. This Workbook is Read by the system using JXL API through Java programming language and stores in a double dimension list.

2) Step 2: Preprocessing - This step takes all the past claims from the database to perform the learning process. And then for each of these claims a list of protocols is being extracted. These protocols are labeled in integers to form the complete integer double dimension list. From this double dimension list required attributes are selected and then they call as the preprocessed list, which is also a double dimension list.

3) Step 3:KNN - In this step the preprocessed double dimension list is taken into account and then for each of the rows of this preprocessed list, all its attributes are summed to find the mean of all attributes for that row and then appended at the end of the row as shown in the Equation 1.

$$\mu = \frac{(\sum_{i=1}^{n} xi)}{n} \qquad (1)$$

Where,

xi=Each Attributes

n = number of attributes i.e. 8

Then each of the rows are evaluated for the Euclidean distance with respect to the other rows and they refer as the row Distance. Row distance of each row is used to evaluate the minimum and the maximum distance of the preprocessed list. Once they are evaluated, then this minimum and maximum distance is used to create the boundaries for the required number of clusters using the following algorithm1. Based on these boundaries nearest neighbor clusters are formed.

Algorithm 1: Cluster Boundary formation

// Input : Min_D, Max_D,K
[Min_D: Minimum Distance, Max_D: Maximum Distance, K : Number of Clusters]
// Output : B_{SET} [Boundary List]
Function : boundaryFormation(Min_D, Max_D,K)
Step 0: Start
Step 1: DIST= (Max_D- Min_D) / K
Step 2: ***for*** j=0 **to** K
Step 3: R1= Min_D
Step 4: R2= R1+DIST
Step 5: T_{SET}=Ø
Step 6: $T_{SET\,[0]}$= R1 , $T_{SET\,[1]}$= R2
Step 7: ADD T_{SET} TO B_{SET}
Step 8: R1= R2
Step 9: **End *for***
Step 10: return B_{SET}
Step 11: Stop

4) Step 4: Regression Analysis and Entropy Estimation - The Input claims list formed in the step 1 for the doctor's claims is considered in this step. Each of the rows from this double dimension list is taken into consideration to find the mean of all the attributes as mentioned in equation 1. Each of this attribute mean is estimated with each cluster's row for the same attribute mean factor with an absolute difference of less than or equal to 0.1.

Each row of the cluster is being counted for this kind of rows and then they are subject to find the entropy of the whole cluster using the Shannon information gain theory as mentioned in equation 2. The cluster with the biggest gain factor is selected for the each of the input claim ID.

$$E = -\frac{X}{Z}\log\frac{X}{Z} - \frac{Y}{Z}\log\frac{Y}{Z} \qquad (2)$$

Where

X= Row Count

Z= Total number of rows of a cluster

Y= Z-X

E = Entropy Gain factor

5) Step 5: Correlation Estimation - Each of the claim index is bonded with a cluster as seen in the last step. Then these input claim indices are now counted for the respective disease in the bonded clusters to form a double dimension list that contains claim index and count of the disease in the bonded cluster and it is referred as the correlation list.

6) Step 6: ANN and Fraud Claim estimation - Each of the correlation list index is evaluated for the cluster rows. This evaluation is carried out by estimating the attribute mean as mentioned in equation 1 with each neuron

cluster's row for the same attribute mean factor of the claim index for the value greater than 0.4. And this is counted to form the neuron cluster count with respect to the given claim index or ID.

Then the correlation list count is used to summing or differencing the neuron cluster count to estimate the fraud level. Which is then displayed to the Health insurance staff who was fed the claim data to the system in the initial.

IV. RESULT & DISCUSSIONS

The Proposed methodology of health care fraud detection for the claims of the Doctors is deployed using the Java Programming Language. The model uses Netbeans 8.0 as the Integrated Development Environment and Mysql as the database server. This software is developed in Windows machine with Core i5 Processor with Primary memory of 6GB. Some experiments are conducted to measure the effectiveness of the Health care fraud claims identification as described below.

Percentage of Error using Mean Absolute Error (MAE)- To measure the percentage of error that proposed system may commit MAE is used for the same. As this measuring parameter provides a better way to estimate the performance of the proposed model. MAE is a measure of difference between the two continues entities which in turn represents the same phenomenon.

Here in this evaluation the percentage error is measured for the estimated claim detected. In this experiment an absolute difference between the Actual Fraud claims and the detected fraud claims are used to evaluate the MAE based on the equation 3.

$$\text{MAE} = \frac{(\sum_{i=1}^{n} | xi - yi |)}{n} \qquad (3)$$

Where,

xi - Number of Actual Fraud Claims existed
yi - Number of Detected Fraud Claims
n- Number of Trails

No of Input Claims	Actual No of Fraud Claims (xi)	Detetected No of Fraud Claims (yi)	xi-yi
5	2	2	0
10	3	3	0
15	5	5	0
20	6	5	1
25	8	6	2
30	11	11	0
35	12	12	0
40	14	12	2
45	15	12	3
50	16	11	5
		MAE	**0.13**

Table 1: MAE measurement Data

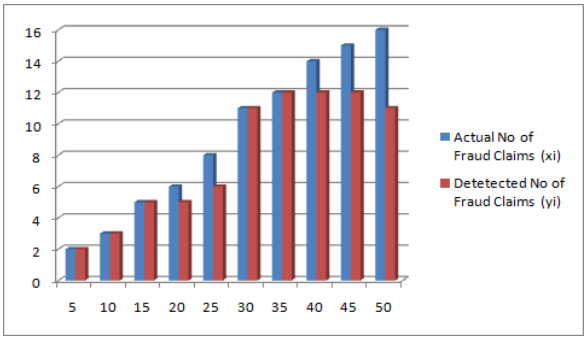

Fig. 2: MAE Evaluation

Table 1 shows some facts of the calculation process that is followed to estimate the Mean absolute error, the effect can be seen in figure 2. The Proposed model yields a MAE of around 0.13 that is too low. And it is the better result of the proposed model in the first attempt only.

V. CONCLUSION & FUTURE SCOPE

The proposed model for Fraud claim identification in health insurance segment uses the Machine learning technique extensively. Here this model uses the KNN clustering algorithm to bring all possible claims that are having a proneness towards the fraud claims. And again, these clusters are evaluated for the closest fraud possibility using Logistic Regression, Entropy Analysis and correlation technique. Then ANN is enhanced to scheme out the fraud claims efficiently when it is blended with Correlation results. And also experiments are indicating that the proposed model achieves a MAE of 0.13, that's really a strong result in the Fraud claim detection process.

In the future this model can be enhanced to work on other complex attributes of the medical field for real time implementation .The result of fraud claims can be brought into the public domain to unleash the Doctors who follow this unethical practice using a web portal.

REFERENCES

[1] M. Farhadi, H. Haddad and H. Shahriar, "Stattic analysis of HIPPA Security Requirements in Electronic Health Record Applications", 42nd IEEE International Conference on Computer Software & Applications, 2018.
[2] Musheer Ahmed and MustaqueAhamad, "Combating Abuse of Health Data in the Age ofeHealth Exchange", IEEE International Conference on Healthcare Informatics, 2014.
[3] Richard A. Bauder and Taghi M. Khoshgoftaar, "Medicare Fraud Detection using Machine Learning Methods", 16th IEEE International Conference on Machine Learning and Applications, 2017.
[4] K.S. Ng, Y. Shan, D.W. Murray, A. Sutinen, B. Schwarz, D. Jeacocke and J. Farrugia, "Detecting Non-compliant Consumers in Spatio-Temporal Health Data:A Case Study from Medicare Australia", IEEE International Conference on Data Mining Workshops, 2010.
[5] P. Bharathi, K.Ramalinga Reddy and G.Srilakshmi, "Medical Image Retrieval Based on LBP HistogramFourier Features and KNN Classifier", IEEEInternational Conference on Advances in Engineering & Technology Research, 2014.
[6] Jawad H AIKhateeb, Fouad Khelifi, Jianmin Jiang and Stan S Ipson, "A New Approach for Off-Line Handwritten ArabicWord Recognition Using KNN Classifier", IEEE International Conference on Signal and Image Processing Applications, 2009.
[7] Ines Ben Fredj and KaïsOuni, "Comparison of Crisp and Fuzzy kNN in PhonemeRecognition", International Conference on Advanced Systems and Electric Technologies, 2017.
[8] Tao Dong, Weinan Cheng and Wenqian Shang, "The Research of kNN Text Categorization Algorithm Based

on Eager Learning", International Conference on Industrial Control and Electronics Engineering, 2012.
[9] Xiping Wang and Wenxue Tan, "A Survey on Intelligent Information ProcessingSystem: A Machine Ailment Diagnosing Based on KNN Similarity Degree", International Conference on Computer Sciences and Applications, 2013.
[10] Ali Muhtar, I WayanMustika andSuharyanto, "The Comparison of ANN-BP and ANN-PSO asLearning Algorithm to Track MPP in PV System", 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 2017.
[11] Reddi Kamesh and Kalipatnapu Yamuna Rani, "Novel Formulation of Adaptive MPC as EKF UsingANN Model: Multiproduct Semi-batchPolymerization Reactor Case Study", Ieee Transactions on Neural Networks and Learning Systems, 2016.
[12] Miguel Antonio Sovierzoski and Fernanda Isabel Marques Argoud, "Evaluation of ANN Classifiers During Supervised Training withROC Analysis and Cross Validation", International Conference on BioMedical Engineering and Informatics, 2008.
[13] Ming-yang Liu, Peng Zhou, Ping Kong, Chun-guang Yang, Gang Li and Ming-ren Mu, "Determination of Octane Number of Gasoline by Double ANN Algorithm Combined with Multidimensional Gas Chromatography", Sixth International Conference on Natural Computation, 2010.

