



e-ISSN: 2319-8753 | p-ISSN: 2320-6710

IJRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.512

9940 572 462

6381 907 438

ijirset@gmail.com

www.ijirset.com



A Model Approach to discover the Health care Fraud Claims

Sayali Bharat Tambe, Sneha Dilip Nalawade, Dattatrya Ramesh Kumbhar, Abhijit Subhash Chatur,
Prof K.S.Avhad

B.E Student, Department of Computer Engineering, SAE, Kondhwa, Pune, India

Professor, Department of Computer Engineering, SAE, Kondhwa, Pune, 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 Decision Tree.

KEYWORDS: K-Nearest Neighbor, Entropy Estimation, ANN, Decision Tree.

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 utilized 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 neighboring 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 neighbor 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 neighbor 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 neighbors, in regression the average value is selected from the collection of k number of neighbors.

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 are the K nearest neighbour doesn't 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 utilized 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 neighboring elements.

Artificial Neural Networks are one of the most essential tools of machine learning. It is widely used in various applications, as it is quite a powerful data modeling tool. The Artificial Neural network is used to find patterns and complex relationships between the values in a dataset, to the 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 modeled 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 is 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 reaches a threshold value.

The Artificial neural network if is setup properly, emulates human-like behavior. 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 analyzing medical images, and they are also quite useful for predicting weather conditions over a certain area given the changing variables that can be utilized by the network to provide valuable insight for the weather prediction.

In this research article related works are mentioned in the section 2. The proposed technique is deeply narrated in the section 3. The experimental evaluation is performed in section 4 and whereas section 5 concludes this research article with the scope for 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 is 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 than the 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 depend 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 method implemented 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.

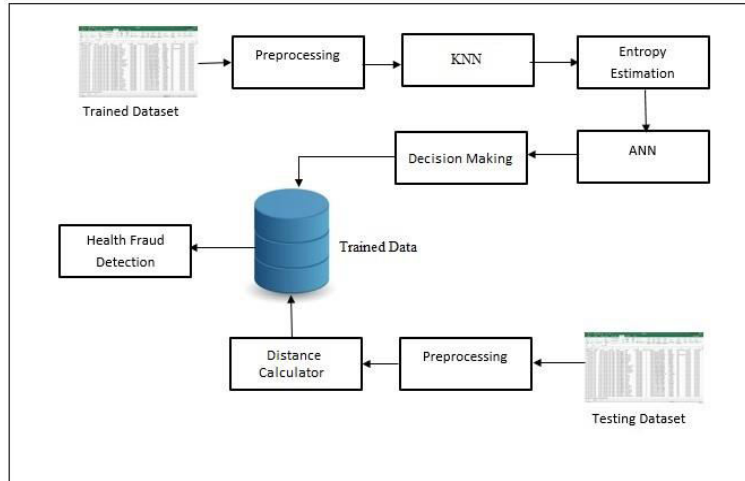
Ali Muhtar [10] Artificial Neural Network is one of the most used technique in the recent years. ANN is for analyzing, learning, modeling, which is the one of the most complicated phenomenon. In this paper their comparison between ANN using PSO and particle swarm Optimization to propagate the maximum 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, mills, 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

Figure 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.

Step 1: Data Collection - The proposed model is designed for the entities like health Insurance company Staff - Receives the claims of many patient that need to be processed for that instance. Therefore, a dataset containing various health insurance claims have been downloaded from the URL - https://www.kaggle.com/rohitro/healthcare-provider-fraud-detection-analysis?select=Test_Inpatientdata-1542969243754.csv

The dataset downloaded from this location contains a workbook file format with the various parameters and attributes related to a particular claim. These attributes contain a lot of different data such as beneficiary ID, Date of Birth, Gender, Race, Renal Disease indicator, state code, country code, chronic disease presence, such as, Alzheimer's, Heart Failure, Kidney Disease, cancer, obstructive pulmonary disease, Diabetes, ischemic Heart Disease, Osteoporosis, Rheumatoid Arthritis, and stroke. The dataset also includes attributes such as IPAnnualReimbursementAmt, IPAnnualDeductibleAmt, OPAnnualReimbursementAmt, and OPAnnualDeductibleAmt which are insurance specific attributes.

As the input to the system which receives the claims in the form of a dataset which is divided into two different parts, training and testing dataset. These datasets 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.

Step 2: Preprocessing – This step takes all the training dataset to perform the learning process. From dataset double dimension list required attributes are selected and then they call as the preprocessed list, which is also a double dimension list. This list only contains the last four columns as they contain the relevant data for the fraud detection and the rest are not useful hence are eliminated.

This step also evaluates the distance between the last four columns of the dataset. These are the only columns that are useful in determination of the health care fraud and have been the main columns of interest for our methodology. These rows are the IPAnnualReimbursementAmt, IPAnnualDeductibleAmt, OPAnnualReimbursementAmt, and OPAnnualDeductibleAmt. The IP reimbursement and Deductible amounts are used to calculate the distance between them and the OP reimbursement and Deductible amounts are used to calculate another distance.



Step 3: K-Nearest Neighbor clustering – 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 respective row as shown in the Equation 1.

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

Where,

xi=Each Attributes

n = number of attributes i.e. 8

Then each of the rows is 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. The resultant cluster list is achieved with 3 clusters of the data and another cluster containing the outliers is generated.

Algorithm 1: Cluster Boundary formation

```
// Input :MinD, MaxD, K
[ MinD: Minimum Distance, MaxD: Maximum Distance, K : Number of Clusters ]
// Output : BSET [ Boundary List ]
Function :boundaryFormation(MinD, MaxD, K)
Step 0: Start
Step 1: DIST= (MaxD. MinD) / K
Step 2: for j=0 to K
Step 3: R1= MinD
Step 4: R2= R1+DIST
Step 5: TSET=∅
Step 6: TSET [0]= R1 , TSET [1]= R2
Step 7: ADD TSET TO BSET
Step 8: R1= R2
Step 9: End for
Step 10: return BSET
Step 11: Stop
```

Step 4: Entropy Estimation – The clusters achieved in the previous step are considered as an input to this step. The entries of each of the row are evaluated in terms of the attributes ipr and ipd which is the reimbursement and deductible for Inpatient and opr and opd which is the reimbursement and deductible for the outpatient. The differences of these values are calculated and the non-zero values of difference are counted for the inpatient and outpatient. This count for the inpatient and outpatient are considered as the values of x and y in the equation 2 given below. The calculated values of information gain of the row in the clusters are then appended to the end of the list. This list is then sorted in the descending order of the entropy.

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

Where

X= inpatient count

Z= outpatient count

Y= Z-X

E = Entropy Gain factor



Step 5: Artificial Neural Network – The information gain list obtained in the previous step is used to achieve the Artificial Neural Networks. The 4 attributes have been used to achieve the 4 hidden layers for the input data. The information gain values and the clusters are provided as an input to this step of the procedure. The hidden layers utilize the ReLU activation function to identify the error probability through the use of random weights, 1 bias weight each for each layer and the use of two target values as 0.01 and 0.99.

$$Error\ Probability = \sum \frac{1}{2} (T_0 - O_L)^2 \dots\dots (3)$$

Where,

T = Target Values

O_L = Output Layer Values

These values are then used in the equation 3 below to estimate the error probability through the use of the values achieved from the output layer. This error probability values are then stored as an error probability list which is sorted in the ascending order of the probability values. The count of this list is then divided by 2 and the resultant top half of the list is considered.

All the achieved data is added to the distance lists which are then sorted in the ascending order. The top ten values from these lists are added to the com list. The unique values from the list are then considered and the smallest value from both the list is then extracted and stored into the database. This ends the training process.

Step 6: Decision Tree – The testing is performed through the testing dataset provided as an input to the system. The excel sheet is read in the form of a double dimension list. The relevant values of ipr, ipd, opr and opd are extracted and the distance between these values is evaluated. The distance values are then subjected to cleansing, where only the non-zero values are selected.

These values are compared with the distance values stored on the database. These are then provided to the Decision Tree to evaluate the presence of any fraud which is indicated by lower values of distance than the distance values achieved in the training phase that are stored in the database. Once the health care fraud in the testing data is detected, it is displayed to the user through the Graphical User Interface.

IV. RESULT AND 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 8.2 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 represent 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.

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

Where,

xi - Number of Actual Fraud Claims existed



y_i - Number of Detected Fraud Claims
 n- Number of Trails

No of Input Claims	Actual No of Fraud Claims (xi)	Detected 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

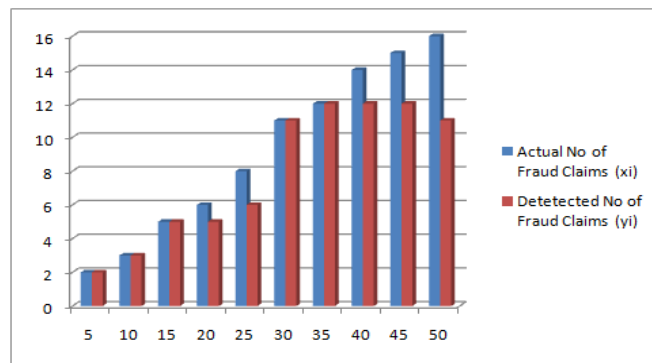


Figure 2: MAE Evaluation

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 AND FUTURE SCOPE

The proposed model for Fraud claim identification in health insurance segment uses the Machine learning technique extensively. The dataset is effectively preprocessed and labeled before providing it to the next step for Cluster formation. Here this model uses the K-Nearest Neighbor clustering algorithm to bring all possible claims that are having proneness towards the fraud claims. Here this model uses the entropy estimation algorithm to achieve the information gain to determine the irregularities in the insurance claim. And again, these entropy values are evaluated for the closest fraud possibility using Entropy Analysis. Then Artificial Neural Network is enhanced to scheme out the fraud claims efficiently when it is blended with entropy results. The Artificial Neural Network approach effectively performs the error probability estimation to achieve accurate and highly insightful analysis of the fraud cases. To determinethe performance, extensive experiments are performed for 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, "Static 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 of Health 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 Histogram Fourier Features and KNN Classifier", IEEE International Conference on Advances in Engineering & Technology Research, 2014.
- [6] Jawad H AIKhateeb, FouadKhelifi, Jianmin Jiang and Stan S Ipson, "A New Approach for Off-Line Handwritten Arabic Word Recognition Using KNN Classifier", IEEE International Conference on Signal and Image Processing Applications, 2009.
- [7] Ines Ben Fredj and Kaïs Ouni, "Comparison of Crisp and Fuzzy kNN in Phoneme Recognition", 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 Processing System: A Machine Ailment Diagnosing Based on KNN Similarity Degree", International Conference on Computer Sciences and Applications, 2013.
- [10] Ali Muhtar, I Wayan Mustika and Suharyanto, "The Comparison of ANN-BP and ANN-PSO as Learning 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 Using ANN Model: Multiproduct Semi-batch Polymerization 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 with ROC 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.



**Impact Factor:
7.512**



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details