ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2015

Forecasting Diseases by Classification and Clustering Techniques

Nikita Ghiya[1], Samruddhi Godbole[2], Pooja Hol[3], Gayatri Deortare[4], Madhuri Chavan[5]

U.G. Student, Department of Computer Engineering, VIIT Engineering College, Pune, Maharashtra, India[1]

U.G. Student, Department of Computer Engineering, VIIT Engineering College, Pune, Maharashtra, India[2]

U.G. Student, Department of Computer Engineering, VIIT Engineering College, Pune, Maharashtra, India[3]

U.G. Student, Department of Computer Engineering, VIIT Engineering College, Pune, Maharashtra, India[4]

Associate Professor, Department of Computer Engineering, VIIT Engineering College, Pune, Maharashtra, India[5]

ABSTRACT: In medical industry there is a huge amount of patients data which is not mined. This healthcare data can be used to extract knowledge for further disease prediction. Currently data mining techniques are widely used in clinical expert systems for prediction of various diseases. These techniques discover the hidden relationships and patterns of the healthcare data.No such expert system which can predict more than one disease exists till date. Almost all other systems use clinical data having parameters and inputs from the tests conducted in laboratory. Very few expert systems are based on the risk factors affecting the disease such as heart disease and diabetes. By using K-means Clustering Algorithm(KCA) in our proposed system, the disease can be predicted more accurately and in less time. Such systems will warn the people about the presence of their disease even before he concerns the doctor. This can even help doctors to carry out specific tests of the patients and target out the disease.

KEYWORDS: Clustering, Fuzzy logic, K-means, Stemming, Stopwords

I. INTRODUCTION

Data mining is a technique to extract knowledge and capture fundamental relationships among the database, extraction of pattern from data, Information discovery. Data mining can be used to extract the medical data from the healthcare industry Electronic patient records further expands the possibilities regarding medical data mining thereby opening the door to a vast source of medical data analysis. Usual database cannot be used for finding relationship and trends among the data. Such techniques play a vital role in extracting data and capturing fundamental relationship among data sets, one such technique is cluster analysis.[5]

Cluster analysis or *clustering* means grouping a set of objects in such a way that objects in the same group (*cluster*) are more similar to each other than to those in other groups (clusters). Cluster analysis is the general task to be solved, not a specific algorithm. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Clustering is a multi objective optimization problem. The algorithm for clustering and parameter settings depend on the individual data set and intended use of the results. Clustering is an iterative process for discovering knowledge ,interactive multi-objective optimization that involves trial and failure rather then automatic process.Frequently modifying data preprocessing and model parameters is necessary in order to achieve the desired properties.

II. LITERATURE SURVEY.

In the paper "Diagnosis of coronary artery disease using the imperialist competitive algorithm",Zahra Mahmoodi and Mohammad Saniee Abadeh [1] states that pre processing the data is the initial step.This step consist of following actions: 1.filing missing values ,removal of outliners,feature selection .Two methods are described in this paper for hndling the pre processing viz. attribute removal and data imputation.

In another paper,"Development of evolutionary Data mining algorithms and their applications to cardiac disease diagnosis", Jenn-Long Liu [3] proposed a system that uses k-means clustering algorithm to partition cardiac disease database into k-cluster .It proposed hybrid method known as GA-KM that combines a genetic algorithm.The database includes 13 attributes, GA-KM determines the optimal weights of attributes and cluster centers of the database.Genetic algorithm is a search algrithm based on Darwinian principle.This paper suggests that GA generates optimal solution.

Fuzzy_implementation_070423.pdf [7] proposes a method for a fuzzy logic in order to tackle the uncertainties associated with heuristic knowledge.The system uses fuzzification process in which for each input and output variable selected, define 2 or more membership functions. And a qualitative category (e.g. low,high,medium.)were defined for each of them.The next step includes rule based defination. It formed IF-THEN rules based upon fuzzification output. Last step consists of Defuzzifcation such that all the outputs of membership functions were converted to respective percentage value (percentage of risk).

III. CONCEPT OF CLUSTER ANALYSIS

There are so many clustering algorithms.,which work on a group of data objects. However, researchers use different cluster models, which implement different algorithms..

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example a hierarchy of clusters embedded in each other. Clustering's can be roughly distinguished as:

- hard clustering: each object belongs to a cluster or not
- soft clustering : each object belongs to each cluster to a certain degree .

IV.TECHNIQUE FOR CLUSTER ANALYSIS

Cluster analysis Technique:-

- K-means algorithm.

K-means Algorithm[6]

Clustering is a method of grouping objects such that members of same group are similar in some way between themselves and dissimilar to other objects. The dataset is partitioned into K groups, such that k is a non negative integer which is assigned before execution. Consider a real time scenario where books in bookshop are clustered in categories(engineering,arts books in educational books).K-means is one of the unsupervised algorithms for clustering.

The functioning of K-means is as below:

1.Center to the clusters are initialized which must be done in an appropriate way since different result issues are caused due to different location. No particular way is described to initialize centers.Often the process of selection is done randomly The better choice of initializing centroids is to chose centers far away from each other. If the value of k is greater than objects then each object is selected as centroid of the cluster else for each data we calculate the distance to all centroid and minimum distance .The object having minimum distance is grouped in the respective cluster

2. The closest cluster to each point is attributed. This is accomplished by computing distance between each data point and cluster centers. Then the data points with minimum distance from the centre are assigned to that cluster resulting in early partitioned data.

3. The new cluster center is recalculated the position of each cluster is set to the mean of all data points belonging to that cluster.

The steps 2 and 3 are repeated till there is no change from one iteration to next iteration. Even though the procedure always terminates k-means does not necessarily find optimal configuration.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centre.

V. DATA PROCESSING

The dataset used for this work is to be collected from various patient's survey.

In this system, the information filled up by the patients will be in a descriptive format which contains different parameters in the form of symptoms faced by patient. pre-processing means removing the other unwanted parameters from the dataset[1]. Data pre-processing includes following processes:

Fig 1: Block diagram of the system.

6.1 Main attribute identification

In this stage, the patient's information will be parsed to separate main attributes such that they will form a set of symptoms. And the rest unwanted information such as patient's name, address, e-mail, contact number will be removed.

6.2 Stop words removal

At this stage, unwanted words from the information will be removed. Stop words may consist of different prepositions, adjectives etc. A list of all stop words available on the internet could be used for this purpose.

6.3 Stemming

At this stage, all the main attributes are stemmed, that is they are converted to it's original format.

6.4 Tokenization

In this stage, all the stemmed words are tokenized as per the different human body parts they are associated with.

6.5 Top word selection

In this phase, All the top words from tokenized set of data will be selected .May be top 10 or top 15 words could be selected as per the available data.

6.6 Word vector formation

At this phase, all the top words are put inside a important word vector .A vector could be an array ,structure ,list etc.

Analyzing data using predictive data mining concepts.In this phase the extracted patterns are processed using predictive data mining algorithms which would lead to a improved approximate prediction. Data analysis could be done with the help of following processes

6.7 Fuzzification

In this phase, with the help of the fuzzy logic all the top words from word vector are rated as per their frequency of occurrence from patient's description.

6.8 K-means algorithm

With the help of this algorithm, all the related symptoms for a certain disease are clustered together or grouped together.

K-means algorithm is explained in section 5 This is the last step of disease identification. The input to K-means is taken from the defuzzifier.
The Block diagram of Disease Diagnosis System is shown in Fig 1. It consists of Data Collection i.e the raw data, data preprocessing and analysis of data is done at the end. These three steps are shown in the figure below.
.

VI. CONCLUSION

There have been several attempts previously made in order to predict disease based upon patient's raw symptom description. We have tried to use this effective concept for the social cause of preventing the loss of life by predicting the possibilities of disease being developed beforehand in a simple, cost effective
and time efficient manner. This process is made faster and more accurate as well as user friendly by means of using different algorithms such as k-means, fuzzy logic etc.

REFERENCES

.
[1] Zehera Mahmoodabadi and Mohammad Saniee Abadeh "*Diagnosis of coronary Artery Disease Using The Imperialistic Competitive Algorithm*", journal of computer science and engineering,vol..8,no.2,june 2014,pp 87-93.
[2] Abhishek Taneja "*Heart Disease Prediction System Using Data Mining Techniques*" oriental journal of computer science & technology, 2013.
[3]Jenn-LongLiu,Yu-Tzu,Chih-Lung Hung, "*Development of evolutionary data mining algorithm and their applications to cardiac disease diagnosis*", World congress on computational intelligence,2012.
[4]NirmalaDevi.M, Appavu alias Balamurugan.S, Swathi U.V, "*An amalgam KNN to predict Diabetes Mellitus*", 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013).
[5].http://en.wikipedia.org/wiki/Data_mining
[6]. Teknomo,Kardi.K-Means Clustering Tutorial.
[7] Fuzzy_implementation_070423.pdf

BIOGRAPHY

Miss N.Ghiya Born and brought up in Maharashtra,India. Currently pursuing Bachelors degree in computer engineering,at VIIT college,Pune,Maharashtra ,India. Year of completion of UG degree-2015 June. An active member of CSI.

Miss P.Hol, Born and brought up in Maharashtra, India. Currently pursuing Bachelors degree in computer engineering,at VIIT college,Pune,Maharashtra ,India. Year of completion of UG degree-2015 June. An active member of CSI.

Miss S.Godbole, Born and brought up in Maharashtra, India. Currently pursuing Bachelors degree in computer engineering,at VIIT college,Pune,Maharashtra ,India. Year of completion of UG degree-2015 June. An active member of CSI.

Miss G.Deotare, Born and brought up in Maharashtra,India. Currently pursuing Bachelors degree in computer engineering,at VIIT college,Pune,Maharashtra ,India. Year of completion of UG degree-2015 June. An active member of CSI.