Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4 – No.9, April 2016 – www.caeaccess.org

Disease Prediction System using Data Mining Hybrid Approach

Rahul Patil
Assistant Professor
Dept. of Computer Engineering,
Pimpri Chinchwad college of Engineerig, Pune

Pavan Chopade
Student
Dept. of Computer Engineering,
Pimpri Chinchwad college of Engineerig, Pune

Abhishek Mishra
Student
Dept. of Computer Engineering,
Pimpri Chinchwad college of Engineerig, Pune

Bhushan Sane
Student
Dept. of Computer Engineering,
Pimpri Chinchwad college of Engineerig, Pune

Yuvraj Sargar
Student
Dept. of Computer Engineering,
Pimpri Chinchwad college of Engineerig, Pune

ABSTRACT
Earlier as well as nowadays also, the doctors are using trial and error approach for predicting the diseases based on clinical investigations available. To predict the diseases is one of the major challenge in past years and today also. There is great need of some system that predicts the diseases early on the basis of available symptoms and patients health. Because of this it will become possible to cure the people from hazardous diseases which may lead the humans to death for e.g. Cancer, AIDS etc. We are a proposing system which is based on combination of different data mining techniques such as clustering, classification etc. that are useful to predict the patient's disease state. The patient's disease states can be find out by formalizing the hypothesis based on test results and symptoms of the patient before recommending treatments for the prevailing diseases. The basic aim of our system is to assist doctors in diagnosing the patient by analyzing his available data and relevant information.

General Terms
Clustering, Classification, Prediction, Data Analysis

Keywords
Naïve Bayes, symptoms, data mining, database, graph based, partitioning, hierarchical

1. INTRODUCTION
As we know there are numerous kinds of diseases are there in environment. Some diseases can be easily get cured but some are there which have bad impact on human body and may lead to death. Diseases like Colds-Flu-Gripe (CFG), Dengue (De), Malaria (M), Cholera (Cl), Leptospirosis (L), Chikungunya (CG), Chicken pox (CP), Diarrhoea (Di) have similar characteristics at early stage which makes clinician job difficult in diagnosing such life-threatening disease. Similarly, in 2009, H1N1 was spreading fast around the world. It is contagious type of disease and it spreads same as seasonal flu. Its symptoms are same as seasonal flu like cough, fever, sore throat, body ache, and headache. So at early stages it seems like a seasonal flu to Doctors but later on when these symptoms lead to more serious complications, including pneumonia and respiratory failure Doctors came to know this is not seasonal flu and something else. Similarly consider Dengue fever. It is transmitted by the bite of an Aides mosquito infected with a dengue virus. The mosquito becomes infected when it bites a person with dengue virus in their blood. It can't be spread directly from one person to another person. Sometimes, symptoms are mild and can be mistaken for those of the flu or another viral infection by the victims or patients. However serious problems develop if this cannot be treated early like damage to lymph and blood vessels, bleeding from nose and gums. Many of these diseases are preventable if detected at early stages. It may be difficult for clinician as well as the person who is suffering to distinguish between these prevailing diseases due to medical symptom similarity at early stages. Today's such variety of diseases makes us to feel the need of implementing the system which can predict the diseases as early as possible based on patient health status. If it is possible to predict the disease at early stage based on available symptoms we can provide respective medicinal treatment to patient who is suffering.

2. RELATED WORK
Data mining is refer to as mining the knowledge from large amount of datasets. It is also referred to as knowledge discovery from data (KDD). There is large amount of data available in the society in various fields and it is helpful to turn such data into useful information and knowledge. The information gained from this data can be used for different purposes in different applications. Data mining is a multidisciplinary field, which include the work areas like database technologies, machine learning, pattern recognition, information retrieval, neural network, artificial intelligence. There are different data mining techniques which can be used for extracting knowledge from large amount of data.

In clustering method like partitioning based method include k-means algorithm which is efficient in processing large data sets and also it works on only numeric shapes. Also there are more techniques which are more efficient than k-means. Like in hierarchical clustering smaller clusters are generated, which may helpful for any knowledge discovery. Clustering forms the cluster which groups the similar type of objects in one cluster.

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4 – No.9, April 2016 – www.caeaccess.org

Data in real world is dirty i.e. the data is incomplete, noisy, and inconsistent. Hence no quality data then no quality results for mining. For having quality data, data mining provides data mining preprocessing techniques which include data cleaning, data transformation, data selection etc. which remove the noisy inconsistent and incomplete data.

Classification is also one of the technique of data mining where a classifier is constructed to predict categorical labels. It also include different techniques like k-nearest neighbor, neural network, decision tree. The aim to use data mining technique is usually enables one to collect, store, access, process and ultimately describe and visualize data sets.

2.1 Clustering and its Techniques

Clustering is a technique in which objects of similar type from large dataset are grouped into one cluster. It is nothing but partitioning a set of data into a set of meaningful sub-classes called cluster.

2.1.1 Partitioning based method

Partitioning algorithm is a non-hierarchical, it construct various partitions and then evaluate them by some criterion. It construct a partition of a database **D** of **N** objects into a set of **K** clusters, where user should predefined the number of cluster (**K**). K-means algorithm comes under partitioning based method. It is one of the mostly used clustering algorithm.

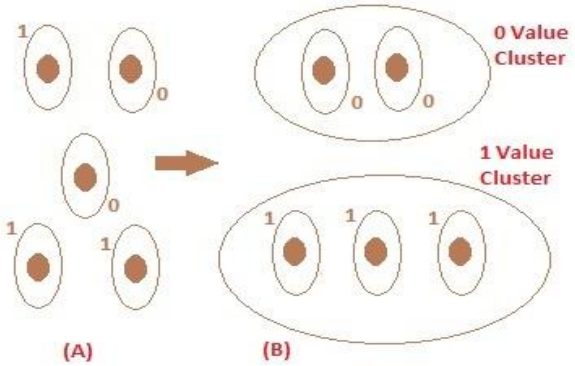


Fig 1: Partitioning based clustering

2.1.2 Hierarchical Clustering

Hierarchical clustering builds a hierarchical decomposition of the set of data or objects using some criterion. It can be visualized as a tree like diagram that records the sequences of merges or splits. Any desired number of cluster can be obtained by 'cutting' the tree at the proper level.

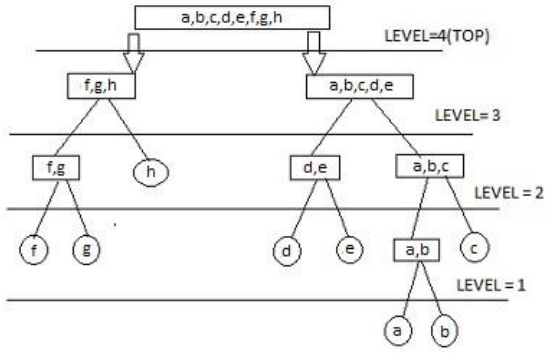


Fig 2: Hierarhical based clustering

2.1.3 Graph based clustering

Graph clustering is the task of grouping the vertices of the graph into clusters taking into consideration the edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters.

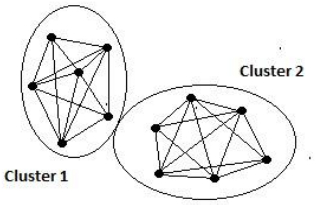


Fig 3: Graph Based Clustering

2.2 Classification

Classification is a data mining machine learning technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks. The Naive Bayesian classifier is also a classification algorithm. It is based on Bayes theorem. A Naive Bayesian algorithm is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

$$P(c/x) = \frac{P(x/c) * P(c)}{P(x)}$$

Fig 4: Naive bayes formula

Where,

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

3. PROPOSED SYSTEM

Earlier as well as nowadays, the doctors are using trial and error approach for predicting the diseases based on clinical investigations available. To predict the diseases is one of the major challenge in past years and today also. There is great need of something that predicts the diseases early on the basis of available symptoms. We are proposing a system which can predict the disease of patients or victims based on his/her symptoms. Data mining techniques like classification, clustering are helpful for prediction purposes. Using data

mining hybrid approach that is using different data mining techniques and algorithms we are predicting the possibilities of having disease and remedies for that. Also the next iteration of the system evolution will propose the report suggestions for the specific diseases. This paper is for such system that can predict the disease state i.e. The person is suffering with disease or not. The working flow for the prediction system is as below.

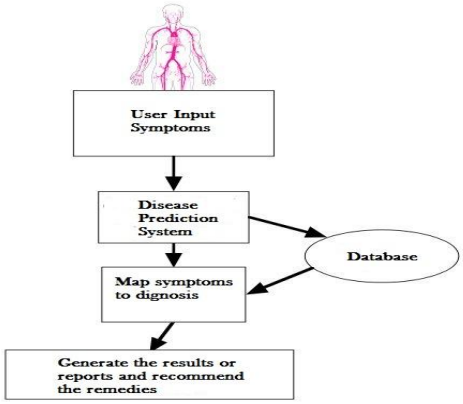


Fig 5: Workflow for disease prediction system

The User will enter their symptoms according to the disease states he is suffering from to the disease prediction system and then the symptoms would get analyzed with the previously entered parameters for the diseases. This mapping of user symptoms and the prior database is once done then the result will generated according to the disease state and level of affection. The detailed system flow will be like as below:

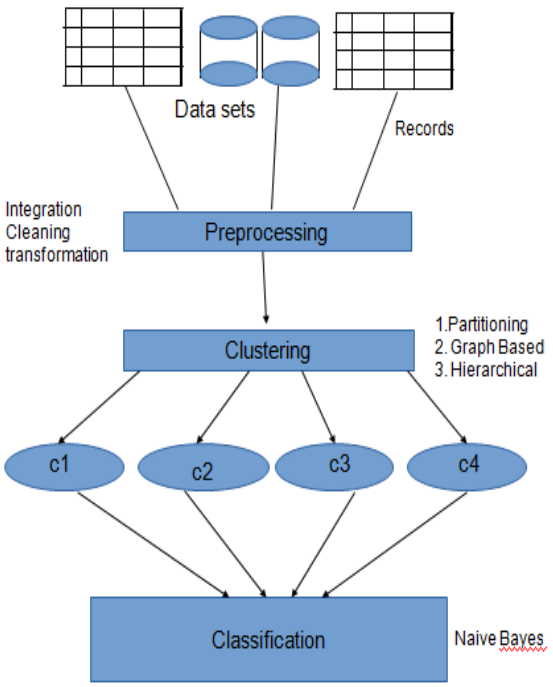
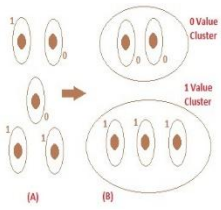
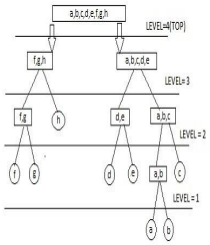
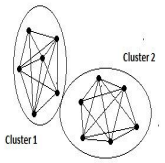


Fig 6: The detailed flow for the disease prediction system.

Table 1. Comparison Between Clustering Techniques

Sr. No	Clustering Techniques		
	Cluster Forms	**Methods**	**Algorithms**
1		Partitioning method.	K-Means
2		Hierarchical Clustering	Agglomerative hierarchical Clustering
3		Graph based clustering	Minimum Spanning tree

4. CONCLUSION

In this paper we are proposing such a system which can predict the diseases based on the input symptoms provided by user and help mankind or society people to analyze and understand their health status. This will also provide remedy for calculated particular diseases. People can self-analyze their health state and can take precautions as per the results. It would help the practitioners / Doctors to analyze the health state of the patient and based on that the manual diagnosis of the disease can also be possible by using the disease prediction system. This will provide early diagnosis of the disease which is not possible with the help of manual diagnosis.

5. REFERENCES

[1] Han and M. Kamber, "Data mining: concepts and techniques," 2nd ed. San Francisco: Morgan Kaufinann, Elsevier Science, 2006.

[2] World Health Organization, Available :http//www.searo.who.intlen/SectionIOI

[3] Sofianita Mutalib, Nor Azlin Ali, Shuzlina Abdul Rahman and Azlinah Mohamed, "An Exploratory Study in Classification Methods for Patients Dataset," International Conference on Data Mining and Optimization, 2009, pp. 79-83.

[4] Shameem A. Fathima and Nisar Hundewale, "Comparitive Analysis of Machine learning Techniques for classification of Arbovirus," International Conference on Biomedical and Health Informatics, Hong Kong and Shenzhen, China,Jan 2012

[5] Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.

[6] Archana L. Rane "Clinical Decision Support Model for Prevailing Diseases to Improve Human Life Survivability" 2015 International Conference on Pervasive Computing (ICPC)

[7] Osama Abu Abbas, "Comparison between data clustering algorithms", The Internantional Arab Journal of Information Technology, vol.5, No.3, July 2008