Disease Prediction System using Fuzzy Logic and K-Means Clustering Algorithm

Gayatri Deotare[1] Nikita Ghiya[2] Samruddhi Godbole[3] Pooja Hol[4] Madhuri Karnik[5]
[1,2,3,4]U.G Student [5]Assistant Professor
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]VIIT Engineering College, Pune, Maharashtra, India

Abstract— In health care ecosystem there is large amount of rich information database available, but they lack inadequate techniques to gain useful information from the database. This is because of improper use of techniques to discover hidden relationship between the data. Useful information can be obtained by applying various data mining techniques on the attributes. The information which is gained can be applied for accurate disease predication and a well as treatment . Various clustering algorithms are used to mine the unsupervised data that is on dataset. K-means clustering algorithm (KCA) gives highest accuracy among the other clustering algorithm. Disease can be predicted at early stages and in less time and only from the symptoms and risk factors of affecting the individual. This paper involves preprocessing techniques, fuzzy logic and KCA which increases the accuracy of our proposed system.

Key words: Data mining, K-means clustering algorithm (KCA), fuzzy logic, preprocessing

I. INTRODUCTION

Data mining refers to extracting or mining knowledge and discovering latent patterns which can be converted into valuable information by digging the raw data. It is a powerful new technology with great potential to help medical institutions focusing on the most important information in their disease databases. Also, data mining is an important step in the Knowledge Discovery in Database (KDD) processes that consist of applying data analysis and knowledge discovery algorithms to produce useful patterns (or rules) over the datasets.

As health care industries produces huge amount of data, data mining techniques can be used on such huge medical data to discover useful patterns that can be used in diagnosis and decision making. To detect a disease various tests are to be done by the patient. But using data mining technique the number of test can be reduced. There are various diseases predicted by using data mining techniques but we are focusing on four most locally frequent disease i.e. heart disease, thyroid, hepatitis, breast cancer.

II. LITERATURE SURVEY

The idea proposed in this paper is based on many of the previous work like Ladda [4] describes that in fuzzy logic rule base is the main part in fuzzy inference system and quality of results in a fuzzy system depends on the fuzzy rules. The goal of fuzzy system is to analyze top word that is only the important words. The basic blocks of fuzzy logic are: preprocessing, top words, fuzzifier, inference engine, defuzzifier, fuzzy rule set.[3] introduces a fuzzy classification systesm (FCS) includes two basic steps (1) knowledge extraction which contains a set of fuzzy rules as extracted knowledge and (2) inference engine which classifies the input samples according to fuzzy rule set and reasoning method

In the paper "Diagnosis of coronary artery disease using the imperialist competitive algorithm", [5] states that preprocessing the data is the initial step. This step consist of following actions: filing missing values, removal of stopwords, stemming.

III. PROPOSED SYSTEM

A. Patient Description:

User would enter the symptom description in textual format. The description would be such that all the unusual changes in the body parts is entered.

B. Preprocessing Stage:

The textual description would go under pre-processing as follows:

1) Stop Words Removing:

From the input string, all the stop words are removed at this stage. Stop words are those words which doesn't possess any meaningful information (e.g. 'the',"my","I","at" etc.). String after stop words removal would be easier for processing, and less complex. A list of frequently used stop words is downloaded from internet and if a stop word is present in the input string it is simply omitted with the help of simple java function.

2) Stemming:

Stemming is the process of converting into original form. All the words from input string are stemmed at this step. We get the "root" word from the original word by stemming. For eg. Vomiting is converted to vomit.

3) Top Word Vector:

At this stage, each word's frequency is calculated and words are arranged in descending order as per it's frequency count. An array list would be maintained containing all the words arranged in descending order.

4) Word Vector Formation:

A vector of important words from top word vector (i.e top 10 words from a top word vector) is formed and stored in an array list. This would be named as "important word vector".

C. Fuzzy Logic;

1) Fuzzy Inference Engine:

Initially, all the words from the important word vector(IWV) are matched with the IWVs of each documents (here, each document contains most commonly found symptom description, and pre-procesing steps are applied to all the documents to obtain important word vector for each disease). Frequency count of each word is "ANDed" with frequency count of words from each document.

An array list is maintained containing AND results of each word from IVW. A max and min value for frequency is calculated and it is divided into 5 equal ranges.

All rights reserved by www.ijsrd.com

These five represent s five classes of classification. All above classes are ranked as : Very high=5, high=4, medium=3, low=2, Very low=1.

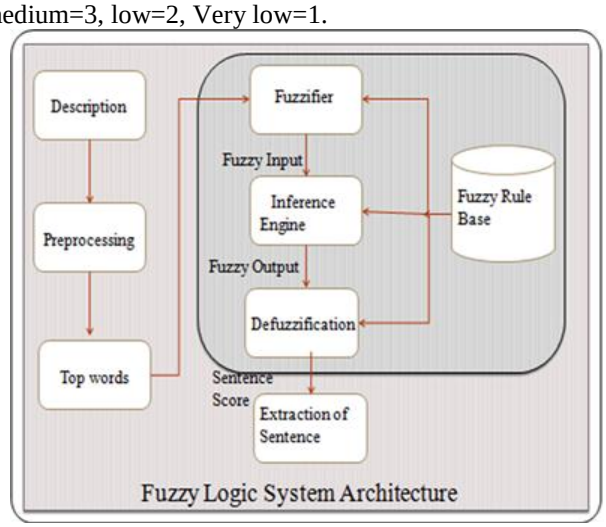

Fig. 1: Fuzzy Logic System Architecture

2) If-Then Rules:

IVWs for each documents are reconsidered and based on ranking values,each words are ranked accordingly. A set of if-then rules is defined as: IF a document has min 1 word with rank 5 AND min 2 words with rank 4 THEN predict disease.

3) Disease Statistics:

A result of IF-THEN rule step is evaluated, and the disease corresponding to that document would be showed as predicted disease.

D. K-Means Clustering:

1) K-Means Clustering and Data Pruning:

For each document, top 2 attributes that affects the disease are considered. Top attributes are found out from internet. KCA is done based on these 2 attributes. A max value and a min value for each cluster is found out. At the end, finned 5 clusters (Very high, high, medium, low, very low) are formed using data pruning.

2) Disease Prediction Step:

Currently entered data is checked for its corresponding cluster. If it belongs to the highest cluster, then the disease is predicted accordingly. This is the output of the system.

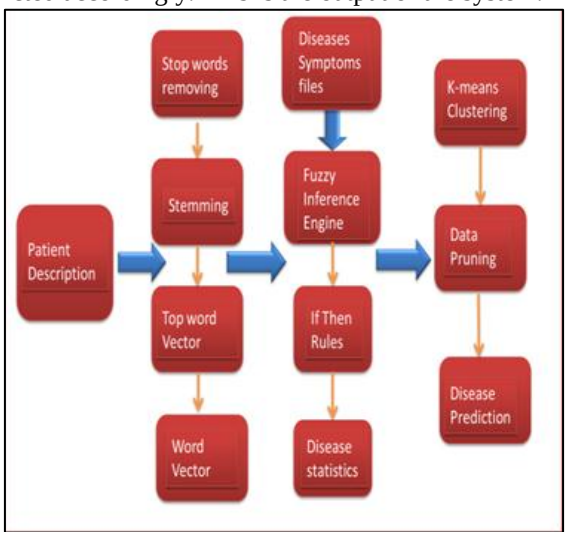

Fig. 2: Block diagram of system

IV. DATA SET

Following is the data set of Heart Disease. This data is taken from UCI repository data set [6].

ID	REFERENCE ID	ATTRIBUTE
1	#3	Age
2	#4	Sex
3	#9	cp: chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
4	#10	trestbps: resting blood pressure (in mm)
5	#12	chol: serum cholestoral in mg/dl
6	#16	fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false
7	#19	restecg: resting electrocardiographic Value 0: normal Value 1: having ST-T wave abnormality Value 2: showing probable
8	#32	thalach: maximum heart rate achieved
9	#38	exang: exercise induced angina (1 = yes; 0 = no)
10	#40	oldpeak = ST depression induced by exercise relative to rest
11	#41	slope: the slope of the peak exercise ST Value 1: upsloping Value 2: flat Value 3: downsloping
12	#44	ca: number of major vessels (0-3) colored by flourosopy
13	#51	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14	#58	Number (The predicted value)

Table 1: Heart Disease data set

This data set is used from data mining by using KCA. Similarly data sets are also available for Thyroid, Breast cancer and Hepatitis on the UCI repository website. We use this data set to predict the disease from the attributes and values provided in that data set. Thus from this data set we can predict the disease of patient more accurately by using KCA.

V. RESULTS AND DISCUSSIONS

To show the effectiveness of proposed system some experiments are conducted on java based windows machine using apache tomcat as server. To measure the performance of the system we set the bench mark by selecting the data set form UCI repository of four diseases like breast cancer, heart disease, hepatitis and thyroid disease as main measuring protocols of disease detection system.

To determine the accuracy of our system, we examined how many relevant diseases are identified based on our fuzzy k-means classification approach.

All rights reserved by *www.ijsrd.com*

To measure this precision and recall are considering as the best measuring techniques. So precision can be defined as the ratio of the number of relevant diseases identified to the total number of irrelevant and relevant diseases identified. It is usually expressed as a percentage. This gives the information about the relative effectiveness of the system.

Whereas Recall is the ratio of the number of relevant diseases are identified to the total number of relevant diseases not identified and it is usually expressed as a percentage. This gives the information about the absolute accuracy of the system.

The advantage of having the two for measures like precision and recall is that one is more important than the other in many circumstances.

For more clarity let we assign

- A = The number of relevant diseases identified,
- B = The number of relevant diseases not identified, and
- C = The number of irrelevant diseases identified.

So, Precision = (A/ (A+ C))*100

And Recall = (A/ (A+ B))*100

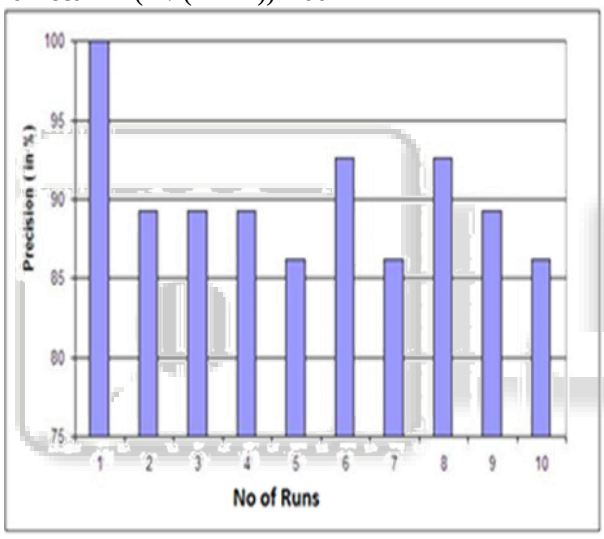

Fig. 3: Average precision of the proposed approach

In Fig. 3, we observe that the tendency of average precision for the identified diseases are high compared to other systems.

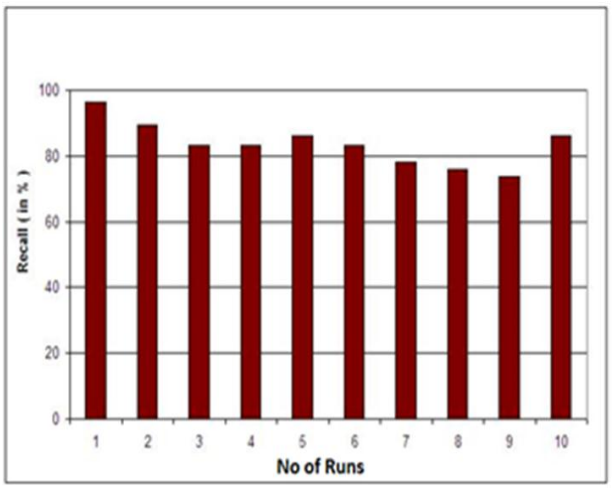

Fig. 4: Average Recall of the proposed approach

In Fig. 4, we observe that the tendency of average Recall for the identified diseases are high compared to other system. So this shows that our proposed system is achieving high accuracy than any other method.

VI. Conclusion

The proposed System predicts the patients disease based on the raw data and description provided by patients. Our System uses fuzzy logic and further K-means Clustering Algorithm (KCA) to predict the disease accurately. Based on historical data, adding attributes information increases the returns while lowering the risks of patients. With our proposed System the disease can be predicted at early stages and by just describing his symptoms. We have use this effective concept of predicting disease as a social cause for preventing loss of life in efficient and user friendly manner. We have given utmost importance to research and thus have formed a dedicated system for prediction of disease.

References

[1] S.Vijiyarani, S.Sudha, "Disease Prediction in Data Mining Technique – A Survey", International Journal of Computer Applications And Information Technology, Volume II, Issue I, January 2013.

[2] Divya Tomar, Sonali Agarwal, "A survey on Data Mining approaches for Healthcare ", International Journal of Bio-Science and Bio-Technology, Vol.5, No.5, 2013.

[3] N. Ghadiri Hedeshiand M. Saniee Abadeh , Coronary Artery Disease Detection Using a Fuzzy-Boosting PSO Approach : Published10April2014

[4] Ladda Suanmali, Naomie Salim, Fuzzy Logic Based Method for Improving Text Summarization , (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009

[5] 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.

[6] D. Aha. (1988, July) Heart disease databases. [Online]. Available: http://repository.seasr. org/Datasets/UCI/arff/heart-c.arff.

[7] Abhishek Taneja "Heart Disease Prediction System Using Data Mining Techniques" oriental journal of computer science & technology, 2013.

[8] 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.

[9] 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)

All rights reserved by *www.ijsrd.com*