ISSN No:-2456-2165

Performance Analysis of Classification and Clustering Based on Mining for Age Related Diseases From Patients Record

J. Jenshya,
M.E Student,
Department of Computer Science and Engineering,
Institute of Road and Transport Technology.
Erode.

DR. A. Kavidha,
M.E., PH.D. Associate Professor.
Department of Computer Science and Engineering,
Institute of Road and Transport Technology.
Erode.

Abstract:- Diseases are the major cause of death and its occurrence increased drastically. Many medical technologies are available for prediction of diseases. In order to analyze, predict and classify the diseases effectively, data mining tools can be used. This work proposes mining of the death causing diseases based on the age with the help of patient's record. For this purpose two different data mining algorithms like classification and clustering are chosen. The two classification algorithm used in this work are decision tree and naïve bayes and the two clustering algorithms that have been chosen are k-means clustering and fuzzy c-means clustering algorithms. The data used in this work is collected from 2000 patient's record which includes the age of the patient and the diseases from which they got affected. The results are obtained in two different ways such as by using classification algorithm and by using clustering algorithm. The performance of classification and clustering algorithms are also analyzed in this work which provides additional information about which algorithm gives the accurate result for this type of data set. The results obtained through this work can be used in various medical fields for disease detection and prevention which may help to minimize the ratio of the death.

I. INTRODUCTION

Data mining is one of the emerging techniques now-a-days. In order to extract the useful information from huge amount of data, data mining technique is used. In other words it can be defined as the process of converting raw data into useful information. Applications of data mining lies in various fields such as in health care systems, educational organizations, business organizations, in detection field of intrusion and lie, financial banking, research analysis and so on. Predictions and classifications are used to discover patterns and relations from the datasets. In data mining the unused data is converted into a dataset for modeling of data by using various techniques of data mining.

The availability of insufficient amount of data in the medical field leads to the increase in death ratio. So it has to be taken consideration in order to minimize the death ratio. In this system the data mining techniques namely classification and clustering is used in order to extract the common diseases related to the age groups with the help of patients record. The two different data mining techniques is used in this system in order to obtain the better results. With the help of these results many medical diagnosis can b carried on in order to predict the age related diseases which will greatly help to minimize the ratio of death. In this project the operation on the datasets were carried out using classification algorithms like Decision trees, Naive Bayes and clustering algorithms like K-means clustering, fuzzy c-means clustering. The prediction can be improved by using association rules to find out the frequently used elements. The implementation first follows by applying the two classification algorithms such as decision tree and naïve bayes method. Decision tree is used to obtain the tree based structure and for the comparison between the ages naïve bayes method is used. So the classification produces the result of the type of diseases based on the age in a tree like manner. The accuracy of the two methods are analyzed.

The second step proceeds by applying the two clustering algorithms such as k-means method and fuzzy c- means clustering method for the data sets. These two methods provides the plot of the diseases based on the age. The ages are provided along the x-axis and the diseases are converted into their corresponding numeric values and those numeric values are provided along the y-axis. The plot indicates which diseases have occurred to the age groups. Finally the performance analysis of the two methods based on the accuracy is also analyzed.

II. SYSTEMANALYSIS

A. Existing system

The existing system had calculated the prevalence of outpatient's diseases through categorization of outpatient's prescriptions and data mining tools were used to identify diseases related to each prescription. It used different classification methods and had compared the performance of those methods with navie method. It used only the prescriptions for calculating the prevalence of outpatient's diseases. The results obtained from this system were based only on the prescriptions and it does not provide any information about the

age related diseases. In another study a range of sources, including electronic patient records, prescriptions and hospitalization data are used for the diagnosis of diseases in individuals. Another study had created a system for prescriptions making that combines association rules and Case Based Reasoning. It used electronic patient records, lab results and symptoms of patients for prescription making. In a similar study Support Vector Machine and Neural Network methods are used for detecting heart disease and prescribing medications for those conditions. That system used the information on patient records, laboratory test results, status of the cardiovascular system, nervous system function, blood pressure, respiratory status and most importantly the information in the Electrocardiogram for diagnosing and prescribing the appropriate medication. In another study Decision tree algorithm is used to determine the stage of type II Diabetes, treatment planning and pharmaceutical dosage.

B. Proposed system

In this project the operation on the datasets werecarried out using classification algorithms like Decision trees, Naive Bayes and clustering algorithms like K-means clustering and fuzzy c-means clustering. The prediction can be improved by using association rules to find out the frequently used.

III. SYSTEM FLOWDIAGRAM

For the implementation of the system the datasets was collected from 2000 patient's records with different ages and different diseases. The datasets is given as the input. The first process proceeds by applying the classification methods such as the decision tree and naïve bayes methods. The results of the two methods are obtained. Along withthis process the performance of the two methods are also analyzed. The second step proceeds by applying the clustering methods such as fuzzy c-means clustering and k-means clustering. The results of the two methods are obtained.Finally the performance of the two methods are analysed.

The two methods are

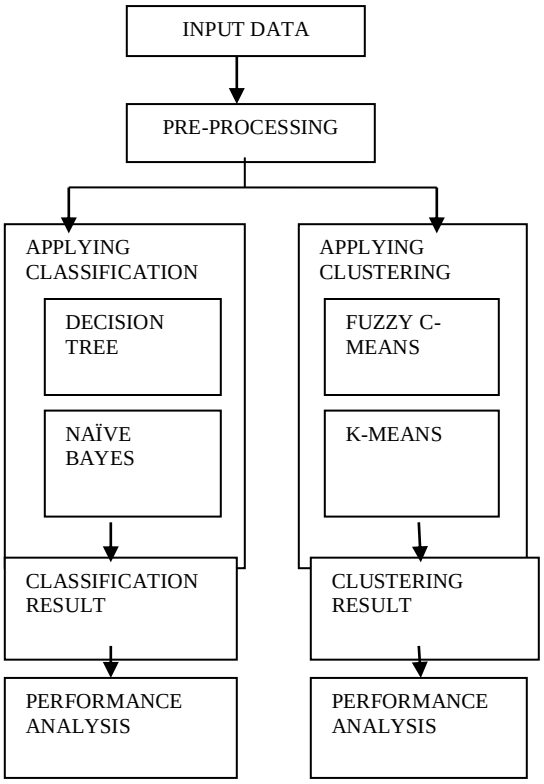
Classification Methods

- Decisiontree
- Naviebayes

Clustering methods

- Fuzzy c-meansclustering
- K-meansclusterin

A. System design



Example decision tree

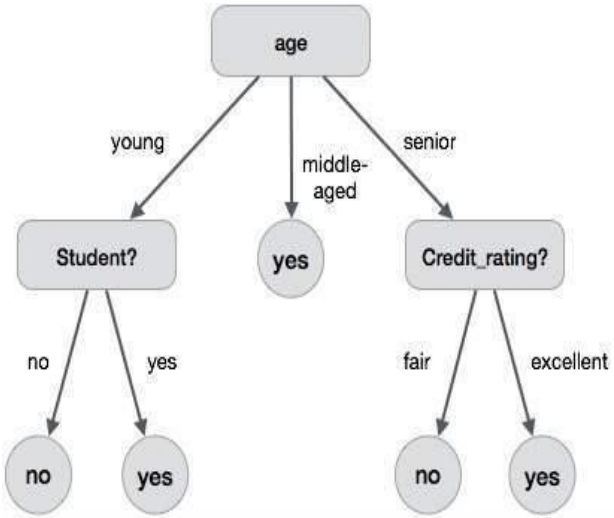


Fig 1:- Example of decision tree

In this project decision tree is used to obtain the tree based structure. Here the decision is made for comparison between the ages. The branches defines the diseases affected to that age groups.

B. Classification methods

➢ *Decision tree*

A decision tree is defined as a graph that uses the branching method in order to illustrate every possible outcome of a decision. It builds classification or regression models in the

form of tree structure. It is easier to understand and allows the addition of new possible scenarios.

➢ *Naïve Bayes*

This model is easy to build and it has no complicated iterative parameter estimation. It is based on Bayes theorem with independence assumption between the predictors. Bayes theorem provides a way of calculating the posterior probability and it is calculated by using the following formula.

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

Where,

P(H) is the probability of hypothesis.
P(E) is the probability of evidence.
P(E | H) is the probability of evidence given that the hypothesis is true.

P(H | E) is the probability of the hypothesis given that the evidence is true.

In this project the Naive Bayes algorithm is used to obtain the conditional probability between the ages and the diseases. Therefore here the decision tree algorithm and the naïve bayes algorithm both are combined to obtain the result effectively.

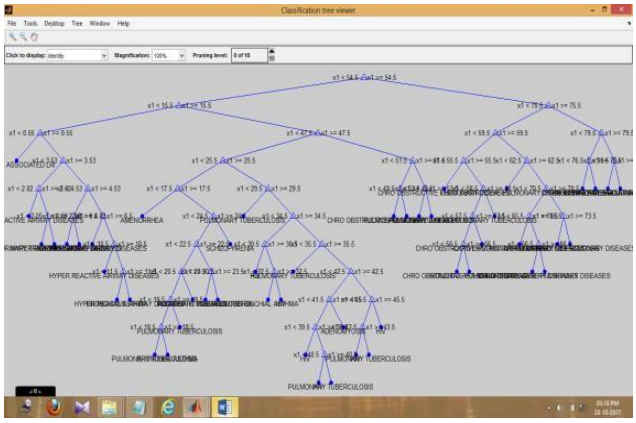


Fig 2:- Result of decision tree and naïve bayes

➢ *Clustering Methods*

- *K-means clustering*

k-means is one of the simplest form of unsupervised learning algorithms that solves the cluster problems. It proceeds as follows

1. Randomly selects the 'c' cluster centers.
2. Calculates the distance between each data points and the cluster centers.
3. Assigns the data points to the cluster centers whose distance from the cluster center is minimum of all other cluster centers.
4. Recalculates the new cluster center by using the following formula:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

5. Recalculates the distance between each data points and the new obtained cluster centers.
6. It stops when no data point was reassigned otherwise it repeats from step 3.

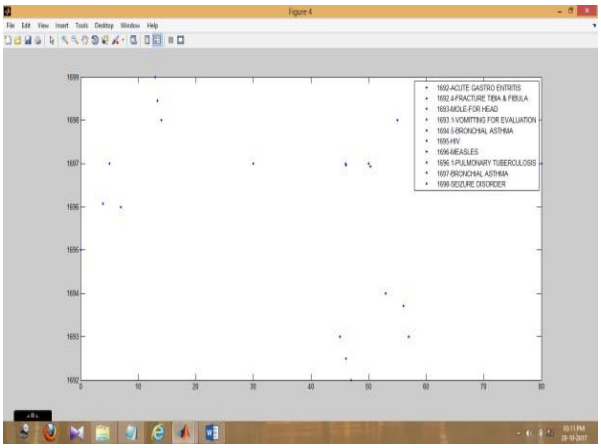


Fig 3:- Result of K-means clustering

- *Fuzzy c-means clustering*

Fuzzy c-means algorithm works by assigning membership to each data point that corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

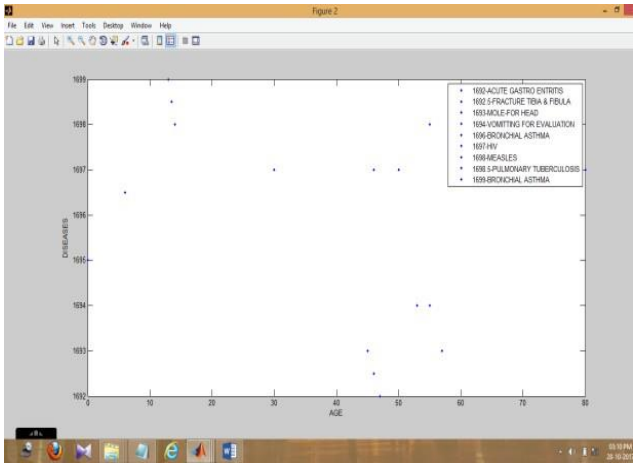


Fig 4:- Result of Fuzzy c-means clustering

IV. CONCLUSION AND FUTURE ENHANCEMENT

A. Conclusion

In this system two different data mining techniques are used to obtain the better results. Performances of those two methods are also analyzed. The results obtained from this system had identified the occurrence of diseases based on the age with the help of patient's records. These results can help to make further diagnosis in medical filed to a larger scale. This clearly shows that the implementation of different data mining tools for this type of system had a better performance when compared with other tools. For this Project the data sets is collected from 2000 patients for various types of diseases from which it is only focused on the identification of diseases based on the age. The two different attributes used in this system are the age of the patients and the disease of the patients. Much medical diagnosis can be achieved through these results. In this system, the existing methods is used to identify the age related diseases from the patient's records and performances of those methods are also analyzed.

B. Future Enhancement

In future it is aimed to implement a new method in order to increase the accuracy and to achieve a better result in the identification of the age related diseases. The performance of the new method is also going to be analyzed. The Future work analyzes the performance of the new method with the performance of the existing method. This will greatly help to obtain the results with higher accuracy and to choose the appropriate method for this type of datasets.

REFERENCES

[1]. S.L. Elo, and I.H. Karlberg , "Validity and utilization of epidemiological data: A study of ischaemic heart disease and coronary risk factors in a local population,"PublicHealth,vol.123,no.1, pp. 52-57,2009.

[2]. J.F. Orueta, R. Nuno-Solinis, M. Mateos, I. Veraqara, G. Grandes, S. Esnaola," Monitoring the prevalence of chronic conditions: which data should we use?," BMC Health Services Research, vol. 12, no. 365, 2012.

[3]. R. Desai, D. Haberling, R.C.Holman,R.J. Singleton, J.E. Cheek, et al., "Impact of Rotavirus Vaccine on Diarrhea-Associated Disease Burden Among American Indian and Alaska Native Children," Pediatrics, vol. 129, no. 4, pp. 907-913, 2012.

[4]. S.L. Ting, W.M. Wang, S.K. Kwork, "RACER: Rule-Associated CasE-based Reasoning for Supporting General Practitioners in Prescription Making," Expert Systems With Applications, vol. 37, no. 12, pp. 8079-8089,2011.

[5]. S.A. Hanna, V.D. Bhagile, R.R. Manza, R.J. Ramteke, "Diagnosis and Medical Prescription of Heart diseases Using Support Vector Machine and FeedforwardBackpropagation Technique," International Journal on ComputreScienceandEngineering,vol.2, no. 6, pp. 2150-2159,2010.