Predictive Data mining Algorithm for Disease Diagnosis

K.Swaathi[1], K.Vidhya[2], V.Vinothini[3], Dr.M.S.Geetha Devasena.(M.E, Phd)[4]
[1, 2, 3, 4] Department of Computer Science and Engineering
[1, 2, 3, 4] Sri Ramakrishna Engineering CollegeCoimbatore

Abstract-*Diagnosis disease are very common and being among the major types of the disease as heart problem, diabetes, coronary artery disease, etc correct and in time diagnosis is very important. Diagnosis method however, it has many side effects and is costly. Existing studies have used several features in collecting data from patients. While applying different data mining algorithms to achieve methods with high accuracy and less side effects and costs. In this paper, a dataset is introduced which utilizes several effective features. Also, a feature creation method is proposed to enrich the dataset. Diagnosis disease most common reason cause death due to coronary artery disease. Then informative gain and confidence were used to determine the effectiveness of the features. Typical chest Pain, breast cancer and age were the most effectiveness ones besides the created features by means of Information gain. Using data mining methods and the feature creation algorithm accuracy is achieved, which is higher than the known approaches in the literature.*

Keywords-Classification, data mining, Diagnosis, coronary artery disease.

I. INTRODUCTION

Data mining is the process of extraction of hidden predictive information from large database is new powerful new technology with great potential to help companies focus on the most information in their datawarehouses. It can reveal the patterns and relationships among large amount of data in a single or several datasets. Data mining is used in various applications such as crime detection, risk evaluation and market analysis. Several industries like banking insurance, and marketing use data mining to reduce costs, and increase profits. Diagnosis disease are among the most common reasons of death all over the world. One major type of these disease is coronary artery disease (CAD). Twenty five percent of people, who have CAD, die suddenly without any previous symptoms. The most important types of disease affecting the heart, and cause severe heart attacks in patients. Being aware of disease symptoms, can aid in time treatment, and reduce the severity of disease's side effects.

Predicting the outcome of a disease is one most interesting and challenging tasks to develop data mining application. The use of computers with automated tools, large volumes of medical data being collected and available in medical research. This kind of difficulty could be resolved with aid of machine learning technique. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment, such as heart and liver disease.

Being expensive and having several side effects, it has motivated many researches to use data mining for diagnosing. Several features and algorithms have been used in the literature. A new feature creation method is used to add three new discriminative features to the patient's records which have a significant impact on prediction ability of the algorithms.

II. EXISTING SYSTEM

Among pre-processing for diagnosis and as far as we know this is the best accuracy so far. Using k-means algorithm on the dataset and reached accuracy. On datasets of low and medium, our algorithm is much faster than other methods, including methods based on low dimensional indexes, such as k-d trees. Other advantages are that it is very simple to implement and it has a very small memory overhead, much smaller than other accelerated algorithm.

III. PROPOSED SYSTEM

The proposed methodology of project deals with evaluation of diagnosis of disease. Various disease like heart problem, diabetes, breast cancer, coronary artery disease, etc., can be proposed diagnosed using data mining techniques. Data mining like classification, clustering, prediction can be used for the diagnosis of different type of disease. Initially, data sets from various data sources are collected, because patient is up to general practitioner who refers the patient.

K -means algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The k -means algorithm is widely used for clustering, compressing, and summarizing vector data. It's used in various applications such as vector quantization, density estimation, and workload

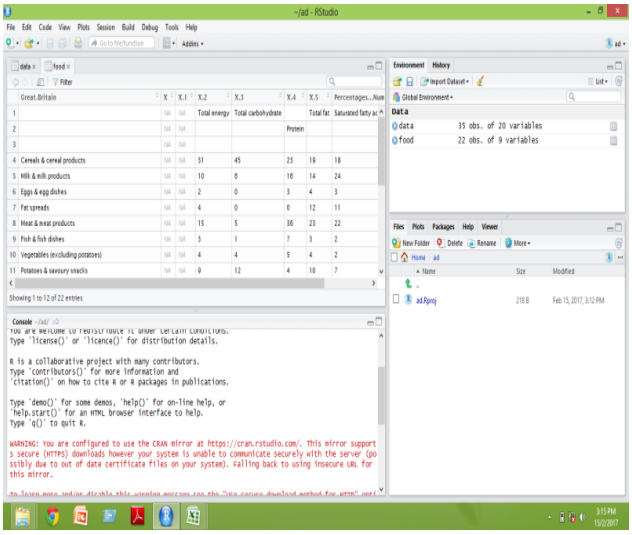
behaviour characterization, among many others. It's been identified as one of the top-10 algorithms in data mining. Because it's such a widely used algorithm, k-means is often used as a subroutine of other learning, coding, and compressing algorithms. The most popular algorithm for k-means is known as Lloyd's algorithm.

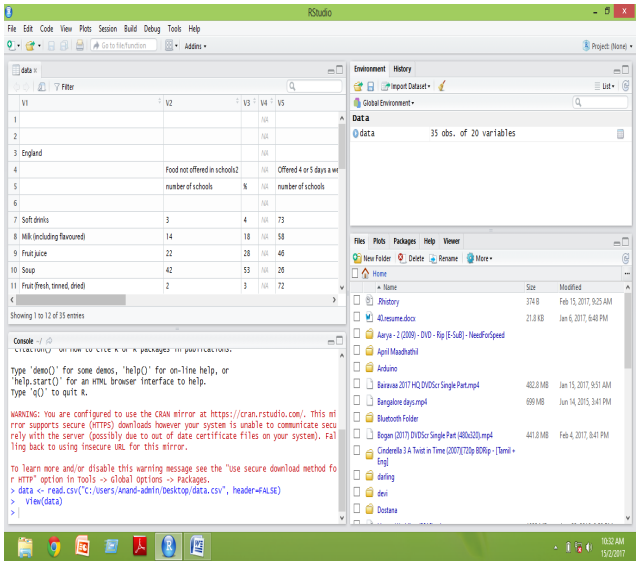
Modules

Module1: Dataset Collection and Preprocessing

- To improve data collection, it should important to know current trends in healthcare.

Purpose of dataset: Identify the data elements that should be collected for each students uniformity in definitions.





Module2: Implementation of k-means algorithm

- The k-means algorithm widely used for compressing, and summarizing vector data.
- We propose a new acceleration for exact k-means that gives the same answer, but is much faster in practice
- Other advantages are that it is very simple to implement and it has a very small memory overhead, much smaller than other accelerated algorithms.

Module3: Performance Analysis

- To analysis the performance of the dataset and classify the accuracy.

IV. FUTURE ASPECTS

The future work regarding diagnosis disease should focus on improving the accuracy by additional features of data mining. In addition, the features used in this study, can be measured with affordable costs and side effects.

V. CONCLUSION

In this study, several algorithms were applied on the dataset and the result were discussed. The features included in this dataset are possible indicators of disease, according to our medical knowledge. In addition, data mining techniques including feature selection and creation were used to improve the accuracy. The accuracy value achieved in this study is, to the best of our knowledge, higher than currently reported values in the literature.

REFERENCES

[1] Roohallah Alizadehasani, Jafar Habibi, Mohammad Javad Hosseini,Hoda Mashayekhi, Reihane Boghrati , Asma Ghandeharioum, Behdad Bahadorian Zahara Alizadeh Sani"A data mining approach for diagnosis artery disease" computer methods and programs in biomedicine biomedicine,2013,Elsevier,vol 111,pp(52-61)

[2] Pinango Dorado "A Bayesian model for disease predictive using symptomatic information" Central America and Convention ,2014,IEEE

[3] Mahmood Hussian Kadhem ,Ahmed M.Zeki" Prediction of Urinary System Disease Diagonsis: A Comparative Study of Three Decision Tree Algorithms " International Conference on Computer Assisted System in Health,2015,IEEE

[4] Jagannatha Reddy,Kavitha"Expert System to Predict the Type of fever Using Data Mining Techniques on Medical Databases" International Journal of Computer Sciences and Engineering,2015,Volume-03

[5] Sumalatha, Muniraj "Survey on medical diagnosis using data mining techniques" International Conference on Optical Imaging Sensor and Security (ICOSS) ,2013,IEEE

[6] Greg Hamerly, "Making k-means even faster" SIAM,2015