ISSN No:-2456-2165

Modernistic Approach to Clustering Algorithms

Ateeq ur Rehman
MCA (IT Infrastructure management services)
Jain Deemed-to-be University
Bangalore, Karnataka, India

Abirami T
Asst. Professor & Project Guide
Jain Deemed-to-be University
Bangalore, Karnataka, India

Abstract :- Cluster is a group of objects that are similar amongst themselves but dissimilar to the objects in other clusters. Identifying meaningful clusters and thereby a structure in a large un-labelled dataset is an important unsupervised data mining task. Technological progress leads to enlarging volumes of data that require clustering. Clustering large datasets is a challenging resource-intensive task and the key to scalability and performance benefits is to use parallel or concurrent clustering algorithms. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Machine learning techniques are widely used in the medical field for the diagnosis of various diseases. A machine learning technique enables us to identify patterns in observed data, build a model according to the observed pattern and predict things without pre-programmed rules. There are two types of machine learning techniques namely supervised and unsupervised machine learning techniques. To diagnose these diseases the data must be clustered before segmentation.
In this project we analyzed Hard C Means (HCM), Fuzzy C Means (FCM) clustering technique and Naive Bayes classification techniques to diagnose the CVD. The medical dataset is analyzed for all the three techniques and the results are evaluated based on metrics like cluster quality, time consumed, true positive rate, false positive rate, precision, recall and accuracy. The results indicated that FCM clustering technique provided a better accuracy of 91% and precision and recall of 91% and 89% respectively in comparison with other two techniques in presence of noise.

Keywords:- *Cluster; Data Mining; Machine Learning; Supervised; Unsupervised; Hard C Means; Fuzzy C Means; Naïve Bayes; Classification; Dataset; Cluster Quality; Time Consumed; True Positive Rate; False Positive Rate; Precision; Recall; Accuracy*

I. INTRODUCTION

Cluster analysis is a study of algorithms and methods of classifying objects. Cluster analysis does not label or tag and assign an object into a pre-existent structure; instead, the objective is to find a valid organization of the existing data and thereby to identify a structure in the data. It is an important tool to explore pattern recognition and artificial learning.

This project involves a complete study, evaluation and investigation of machine learning algorithms and compare their results. In [4] several computational intelligence techniques are studied for heart disease detection, and their study is based in comparing six known classifiers with the data provided by the University of Cleveland. In that research the method with better precision levels was support vector machines (89,49%) surpassing Decision Trees, K-Nearest neighbors and Part. According to [6], diagnosis of cardiac disease it is a relevant issue and many researchers have developed intelligent decision support systems to improve the capacity of medical staff. In that research, a novelty methodology is presented using SAS 9.1.3 software. The results had 89.01% of precision based on the heart disease dataset of the University of Cleveland, and also obtained 80.95% in sensitivity and 95.91% in specificity in the diagnosis of heart conditions. The title of the project is **"Modernistic approach to clustering algorithms".** For the comparison of the clustering algorithms, cardiovascular patient dataset is used.

Cardiovascular diseases are the main cause of death around the world. Every year, more people die from these diseases than from any other cause. According to World Health Organization data, in 2012 more than 17.5 million people died from this cause, and that represents 31% of all deaths registered worldwide. Data mining techniques are widely used for the analysis of diseases, including cardiovascular conditions. It is necessary to use a clustering method for data segmentation according to their diagnosis.

A. Objective

To study, evaluate and investigate the two unsupervised machine learning technique and one supervised machine learning technique with the common dataset as the input. The results of the algorithms are evaluated based on the attributes like cluster quality, time consumed, true positive rate, false positive rate, precision, recall and accuracy. Finally the results are compared and the algorithm which gives better result for the dataset is observed.

B. Scope

This project serves as a baseline to analyze and evaluate different machine learning algorithms for medical dataset, consequently to determine the suitable algorithm to be chosen for a particular data input. This project helps in diagnosing the cardiovascular disease for different machine learning algorithms and evaluates the precision of the diagnosis.

C. Motivation

There are several different algorithms to analyze the input dataset and produce results as clusters or classes. But different algorithms produce variable results for the different input dataset. Because our project revolves around diagnosing the cardiovascular disease whose results are very crucial, the results are required to be precise. Hence we compare different types of machine learning algorithm to learn which algorithm can improve the diagnosis accuracy.

II. ASSUMPTIONS AND DEPENDENCIES

The various assumption and dependencies that affect the project are as follows:

- The project depends on the input dataset and the intended results.
- The results that are obtained from the HCM, FCM and Naive Bayes technique are assumed to be true.
- The user of this project will be concerned to the medical field who wished to know about the persons of being affected by any heart disease symptoms.

III. DATASET ANALYSIS AND PREPARATION

The dataset analyzed for this project is the "Heart disease dataset" which is contained in the Machine Learning Repository UCI [9]. There are 303 records in the dataset and each record consists of 14 attributes. The description of all the attributes of the dataset is given below:

- Age: values of the age of a person in years.
- Sex: male takes the value of 1 and female is 0.
- Chest Pain Type: In case of typical angina the value is1, atypical angina is 2, other kind of pain is 3 and asymptomatic is 4.
- Resting Blood pressure: Value calculated in Hg at the time of hospital admission.
- Cholesterol: mg/dl.
- Blood sugar > 120 mg/dl: 1, in case is true, 0 otherwise.
- Electrocardiogram Result: In case of normal value, is 0, anomaly is 1 and ventricular hypertrophy is 2.
- Maximum heart rate achieved
- Exercise induced angina: In case of negative is 0, otherwise is 1.
- Induced depression.
- Slope peak exercise:
- Number of major vessels (0 - 3) colored by fluoroscopy.
- Thal: in case of normal value is 3, by default is 6, in case of reversible defects is 7.
- Heart disease diagnosis (angiographic disease status): 0 in case of less than 50%, 1 otherwise.

IV. DESIGN

The block diagram specifies the architecture, high level design of a computer program. They aid the programmer in dividing and conquering a large software problem by recursively breaking down the problem into parts that are small enough to be understood by a human brain. The process is called a top-down design, or functional decomposition.

Figure 1 shows the block diagram of the entire system. It shows the decomposition of the entire task and flow from one subtask to another and the interaction between them. It has the following modules.

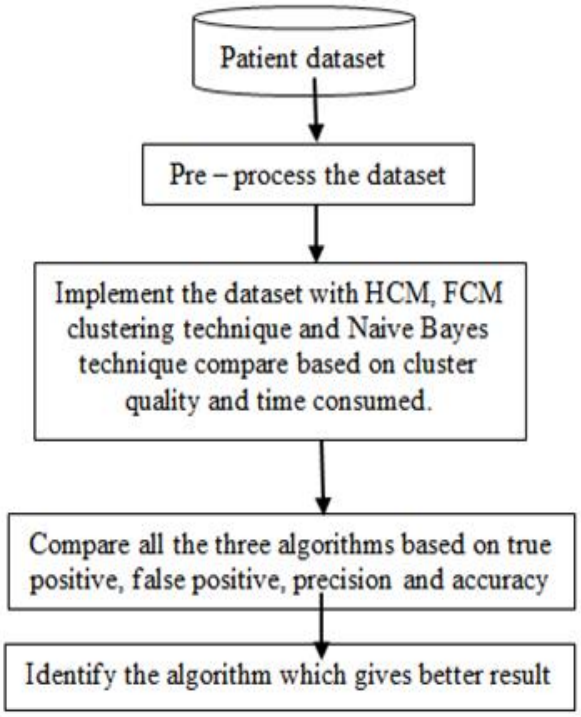

Fig 1

V. IMPLEMENTATION

A. Unsupervised Machine Learning Technique

- *Hard C-Means (K-Means Clustering) Algorithm*

  Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ *be the set of data points and* $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

  Randomly select *'c'* cluster centers.

- Calculate the distance between each data point and cluster centers.
- Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

- Recalculate the new cluster center using: $v_i = \left(\frac{1}{c_i}\right)\sum_{j=1}^{c_i} x_i$ where, c_i represents the number of data points in the i^{th} cluster.
- Recalculate the distance between each data point and new obtained cluster centers.
- If no data point was reassigned then stop, otherwise repeat from step 3).

➢ *Fuzzy C-Mean Algorithm*

Let $X = \{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 ..., v_c\}$ be the set of centers.

- Randomly select *'c'* cluster centers.
- Calculate the fuzzy membership 'u_{ij}' using: $u_{ij} = 1/\sum_{k=1}^{c}(\frac{d_{ij}}{d_{ik}})^{\frac{2}{m}-1}$
- Compute fuzzy centers 'v_j' using : $v_j = (\sum_{i=1}^{n}(u_{ij})^m)/(\sum_{i=1}^{n}(u_{ij})^m)$ for all j=1,2,3.....c
- Repeat step 2) and 3) until minimum j value is achieved or $||u^{(k+1)}-u^{(k)}||<\beta$

  Where, 'k' is the iteration step

  'β' is the termination criterion between [0,1]

  'U=$(u_{ij})_{n*c}$' is the fuzzy membership matrix

  'j' is the objective function

B. Supervised Machine Learning Technique

➢ *Naïve Bayes Classification*

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) * P(x_2|c) * ... ... * P(x_n|c)P(c) \text{ where,}$$

- P(c|x) is the posterior probability of class(c, target) given predictor (x, attributes).
- 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.

VI. RESULT

A. Cluster Quality

The cluster qualities of the proposed techniques are evaluated based on the distance between the clusters and the number of data points that were clustered. The cluster quality should increase as the number of data points in the cluster increases.

Fig 2

B. Time Consumed for Clustering

The time consumed for clustering begins from the time the dataset input for the first iteration of the clustering. The time consumed for clustering should increase as and when the number of records in the dataset increases.

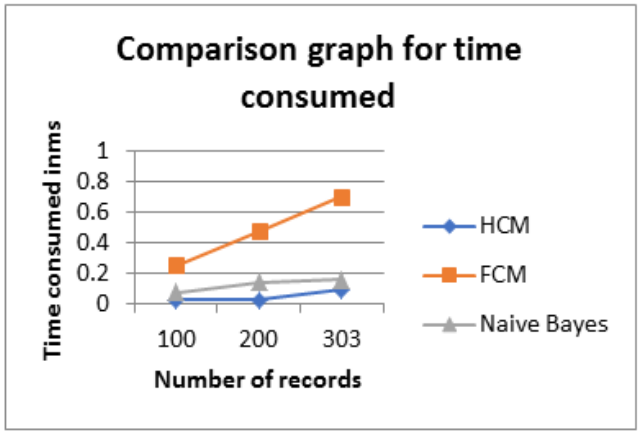

Fig 3

C. Performance

In this section the performance of all the three machine learning techniques are compared based on some more metrics such as true positive rate, false positive rate, precision, recall and accuracy. The metrics such as cluster quality and time consumed are very trivial and do not give accurate results for ambiguous dataset that is used.

➢ *True Positive Rate*

The following is the formula to calculate the true positive rate of the data points.

$$T_p rate = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

➢ *False Positive Rate*

The following is the formula to calculate the false positive rate of the data points.

$$F_p rate = \frac{False\ Positive}{False\ Positive + Verdaderos\ Negative}$$

➢ *Precision*

The following is the formula to calculate the precision of the data points.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

➢ *Recall*

The following is the formula to calculate the recall of the data points.

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + false\ negetivos}$$

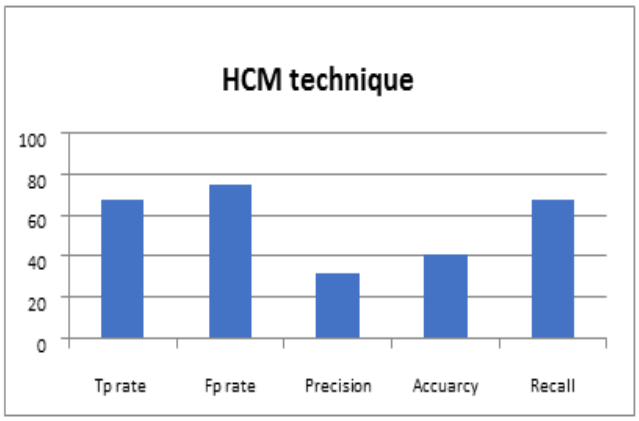

Fig 4

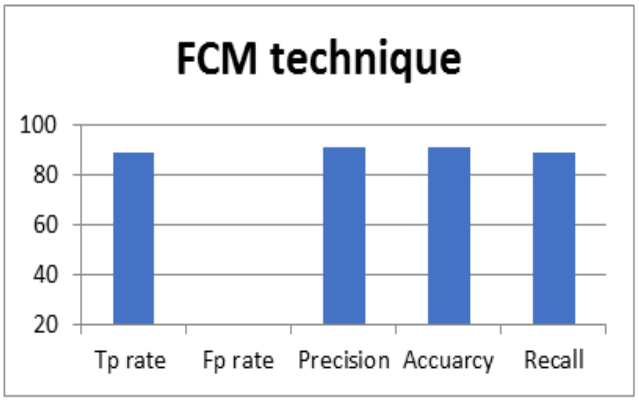

Fig 5

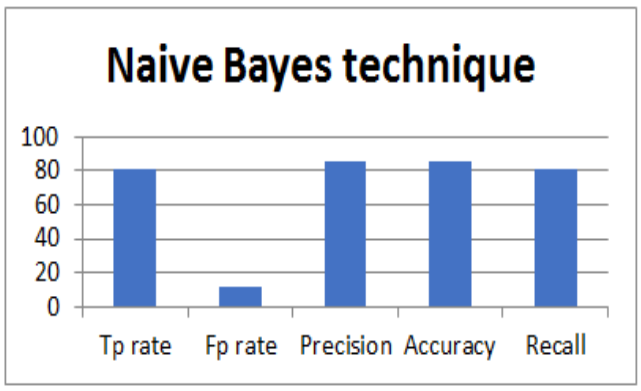

Fig 6

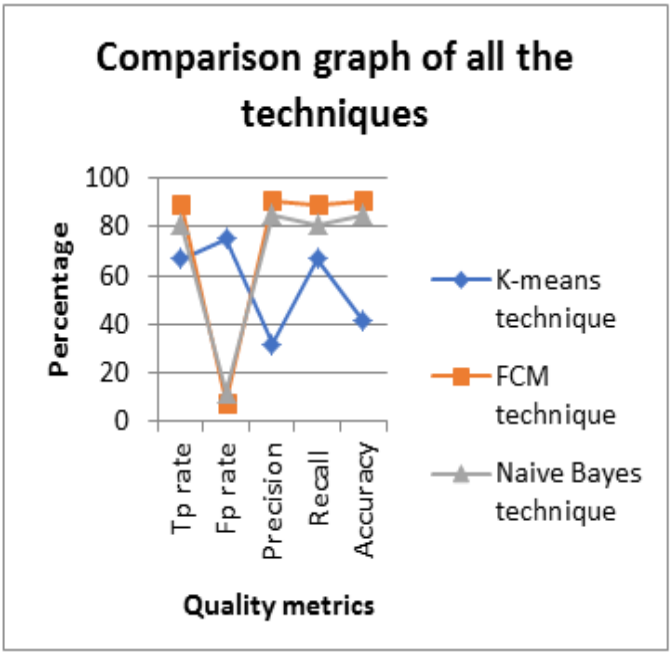

Fig 7

Figure above shows the comparison among the different techniques based on the quality metrics that were used. This figure clearly indicated that without aggregation, if classification technique is compared against the soft clustering technique then soft clustering technique yields the better result.

VII. CONCLUSION

The different machine learning techniques that were used for the implementation are HCM, FCM clustering and the Naive Bayes classification. The clustering techniques HCM and FCM were evaluated based on the cluster quality. All the three techniques were compared for the time consumed for clustering and classification. The time consumption does not stand as a strong base for the evaluation of the clusters. Cluster quality with different metrics are calculated and it was observed that chevbychev distance gives varied results for the same dataset and that metric cannot be used for medical dataset.

We further compare the FCM clustering technique and HCM clustering technique with classification technique Naive Bayes. The techniques were compared based on the true positive rate, false positive rate, precision, recall and accuracy. The results obtained indicate that FCM clustering technique provided a better accuracy of 91% and a precision and recall of 91% and 89% respectively in comparison with the other two techniques in presence of noise. FCM provides better accuracy and precision than the Naive Bayes technique when the data is not aggregated before it is classified.

REFERENCES

[1]. Fabio Mendoza Palechor*, Alexis De la Hoz Manotas, Paola Ariza Colpas , Jorge Sepulveda Ojeda, Roberto Morales Ortega, Marlon Piñeres Melo Universidad de la Costa, Barranquilla, Atlantico, Colombia.(November 2, 2016). Cardiovascular Disease Analysis Using Supervised and Unsupervised Data Mining Techniques

[2]. Burke, A. P., Farb, A., Malcom, G. T., Liang, Y. H., Smialek, J., & Virmani, R. (1997). Coronary risk factors and plaque morphology in men with coronary disease who died suddenly. New England Journal of Medicine.

[3]. El-Hanjouri, M., Alkhaldi, W., Hamdy, N., & Alim, O. A. (2002). Heart diseases diagnosis using HMM. Proceedings of the Electrotechnical Conference on Mediterranean.

[4]. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications.

[5]. Detrano, R., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American. Journal of Cardiology.

[6]. Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications.

[7]. Avci, E., & Turkoglu, I. (2009). An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases. Expert Systems with Applications.

[8]. Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications.

[9]. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. Machine Learning Repository: Fertility Data Set. Recupedao.