International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 7 Issue: 6 29 - 34

SVM Classifier on K-means Clustering Algorithm with Normalization in Data Mining for Prediction

Vasu Deep
Department of Computer Science & Engineering
Translam Institute of Technology & Management, Meerut
Email Id –tovasug@gmail.com

Himanshu Sharma
Department of Computer Science & Engineering
Translam Institute of Technology & Management, Meerut
Email Id – himanshu2210sharma@gmail.com

Abstract-This work is belonging to K-means clustering algorithms classifier is used with this algorithm to classified data and Min Max normalization technique also used is to enhance the results of this work over simply K- Means algorithm. K-means algorithm is a clustering algorithm and basically used for discovering the cluster within a dataset. Here cancer dataset is used for this research work and dataset is classified in two categories – Cancer and Non-Cancer, after execution of the implemented algorithm with SVM and Normalization technique. The initial point selection effects on the results of the algorithm, both in the number of clusters found and their centroids. In this work enhance the k-means clustering algorithm methods are discussed. This technique helps to improve efficiency, accuracy, performance and computational time. Some enhanced variations improve the efficiency and accuracy of algorithm. The main of all methods is to decrees the number of iterations which will less computational time. K-means algorithm in clustering is most popular technique which is widely used technique in data mining. Various enhancements done on K-mean are collected, so by using these enhancements one can build a new proposed algorithm which will be more efficient, accurate and less time consuming than the previous work. More focus of this studies is to decrease the number of iterations which is less time consuming and second one is to gain more accuracy using normalization technique overall belonging to improve time and accuracy than previous studies.

Keywords-*K-mean clustering, Prediction, Normalization, Classification, SVM Classifier.*

*****

I. INTRODUCTION

Data Mining is known as the process of extraction of knowledge or useful patterns from the unorganized and huge data. The goal of data mining is to analyze different type of data by using available data mining tools. The steps in data mining are: Data cleaning, Data Integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. Data mining plays an important role in medical for prediction of disease. There are high risky consequences because of doctor's assumptions and lack of knowledge in particular area. Data mining here plays an important role in discovering or deriving out useful patterns from the historic data of patients, from these patterns prediction analysis for future aspects can be done [1]. Data Mining is the process of analyzing the huge amount of data and encapsulating the relevant information from it. In other words, we can say that data mining is the procedure of mining knowledge from data.

There is very large amount of data everywhere around us. In order to analyze such huge data, powerful approaches ortools arerequired.Such approaches canachieve interesting knowledge fortheusers usingdecision making [2]. Thus, the methodlikedata miningisappliedon implicit,unknown and highly useful data. The process of extracting knowledge from the huge storage databases isknown as data mining.

A. Theoretical Foundations ofData Mining

The theoretical foundations of data mining include the following concepts −

Data Reduction − The basic idea of this theory is to reduce the data representation which trades accuracy for speed in response to the need to obtain quick approximate answers to queries on very large databases. Some of the data reduction techniques are as follows −

- Singular value Decomposition
- Wavelets
- Regression
- Log-linear models
- Histograms
- Clustering
- Sampling
- Construction of Index Trees

Data Compression − The basic idea of this theory is to compress the given data by encoding in terms of the following −

- Bits
- Association Rules
- Decision Trees
- Clusters

Pattern Discovery − The basic idea of this theory is to discover patterns occurring in a database. Following are the areas that contribute to this theory −

- Machine Learning
- Neural Network
- Association Mining
- Sequential Pattern Matching
- Clustering

Probability Theory − This theory is based on statistical theory. The basic idea behind this theory is to discover joint probability distributions of random variables.

Probability Theory − According to this theory, data mining finds the patterns that are interesting only to the extent that they can be used in the decision-making process of some enterprise.

Microeconomic View − As per this theory, a database schema consists of data and patterns that are stored in a database. Therefore, data mining is the task of performing induction on databases.

Inductive databases − Apart from the database-oriented techniques, there are statistical techniques available for data analysis. These techniques can be applied to scientific data and data from economic and social sciences as well.

B. Clustering Methods

Clustering method [3] can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method −Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember −

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods−This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach−This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach−This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages

- The major advantage of this method is fast processing time.

- It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods−In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method−In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

II. REVIEW LITRETAURE

An Algorithm For Predictive Data Mining Approach In Medical Diagnosis, Shakuntala Jatav et.al (February 2018) - In this Paper [4] The Healthcare industry contains big and complex data that may be required in order to discover fascinating pattern of diseases & makes effective decisions with the help of different machine learning techniques. Advanced data mining techniques are used to discover knowledge in database and for medical research. This paper has analyzed prediction systems for Diabetes, Kidney and Liver disease using a greater number of input attributes. The data mining classification techniques, namely Support Vector Machine (SVM) and Random Forest (RF) are analyzed on Diabetes, Kidney and Liver disease database. The performance of these techniques is compared, based on precision, recall, accuracy, measure as well as time. As a result of study, the proposed algorithm is designed using SVM and RF algorithm and the experimental result shows the accuracy of 99.35%, 99.37 and 99.14 on diabetes, kidney and liver disease respectively.

This research paper is mainly focused to predict disease possibility using data mining or machine learning approach in order to enhance the accuracy or precision of the disease detection expert system. This paper also shows the related work study of different approaches such as neural network, naïve bayes, SVM, KNN, FCN, etc and it is concluded that SVM gives the best performance as compared to the other existing techniques. As a result of study, the proposed algorithm is designed using SVM and RF algorithm and the experimental result shows the accuracy of 99.35%, 99.37 and 99.14 on diabetes, kidney and liver disease respectively. In future using data mining approach a new optimized intelligent system can be designed which can give accurate and efficient result.

A Review Paper On Prediction Analysis: Predicting Student Result On The Basis Of Past Result ,PriyankaDhamija et .al, (May 2017) In this paper [6] In today's world competition is increasing day by day. In field of higher education as competition is increasing so the student self-harm rate is increasing. The reason for this is because students are not able to cope with studies and under pressure, they do self-harm. Data mining is a technique which can be used to decrease this self-harm rate. In this paper we want to explain that using data mining we can predict result of students beforehand by using previous year result or any other factors in early stages of any course. This technique is called Prediction Analysis and this an example of use of data mining in the field of education. It can also be called as educational data mining. Any Data mining can be used for this prediction analysis and we will be using WEKA tool of data mining to predict result.

We are trying to predict result of students at early stages only of any course on the basis of previous result using prediction analysis of Data mining. For prediction we are using WEKA tool which is free and is machine learning tool. In WEKA tool we will as input give the past result set on the basis of which a model will be created which provide a particular patter of student result needed to pass exam. The past result set can be made on the basis of any attribute or property of students either their marks or behavior in class. Later on to predict result at any stage of course test set based on same attribute will be made and tested on the model created using WEKA. By doing this institute can help those students before final exam whose result has been predicted fail or low marks.

III. BASIC THEORY

K-mean clustering algorithm: K-means clustering algorithm [10] is one of the most popular and simplest algorithms. It is unsupervised learning algorithm that is used to solve the sound known clustering problems. Procedure followed by it a very simple and easy way to classify a given data set. K-mean clustering algorithm has some properties that are specified below:

- There should be always k cluster.
- Each cluster always contains at least one item.
- Non-hierarchical clusters are formed and they do not overlie

It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to eachother butdissimilar to other objects in other clusters. The algorithm attempts to find the cluster centres, (C1 …… Ck), such that the sum of the squared distances of each data

point, xi ,1 ≤ i ≤ n, to its nearest cluster centre Cj, 1 ≤ j ≤ k, is minimized. First, the algorithm randomly selects the kobjects, each of which initially represents a cluster mean or centre. Then, each object xi in the data set is assigned to the nearest cluster centre i.e. to the most similar centre [5][7].

V. PROPOSED METHODOLOGY

The k-mean clustering algorithm is used to cluster the similar type of data for prediction analysis. In k-mean clustering algorithm, probability of the most relevant function is calculated and using Euclidian distance formula the functions are clustered. In this work, we will enhance the Euclidian distance formula to increase the cluster quality. The enhancement will be based on normalization. In the enhancement two new features will be added. The first point is to calculate normal distance metrics on the basis of normalization. In second point the functions will be clustered on the basis of majority voting. The proposed technique will be implemented in MATLAB.

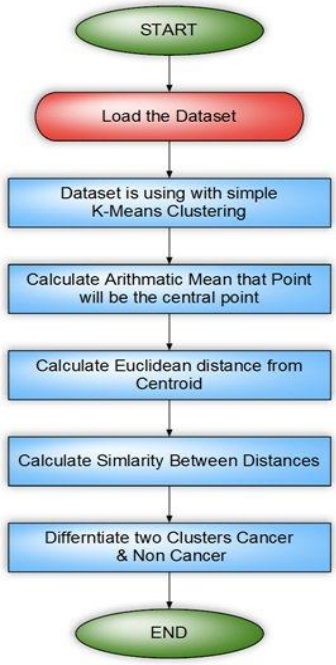

Figure 1: Flowchart of Methodology Working

- First of all, we have started process in which at initial stage we load the dataset from user end.
- When once data has been uploaded successfully then Simple k-means applied and got result in subplot.
- After that calculate the arithmetic mean that point will be the central point.
- Now we calculate the Euclidean distance from the centroid to find out the similarities between distances.
- Now we apply normalization, in precede in which we read text file data of that uploaded data after that we find out the nearest point with normalization to get the best result in which we got result in different form rather than first subplot.
- After this process normalization on that process is done in which iterations process started.
- This process is continued until we don't get a nearest point to accurate position with uploaded data
- After iteration are run successfully then we apply Support Vector Machine (SVM) to classify the data into two clusters Cancer and Non-Cancer.
- At last calculated their total time in which we got results which shows betterment in accuracy of cluster.

VI. EXPERIMENTAL RESULTS

The Ovarian Cancer Dataset is used for this research work and to get the results. The real datasetwas highly dimensional, but only five fields has been finally selected for this research work on the basis of requirements. The dataset is loaded using MATLAB and perform the both algorithms to get results

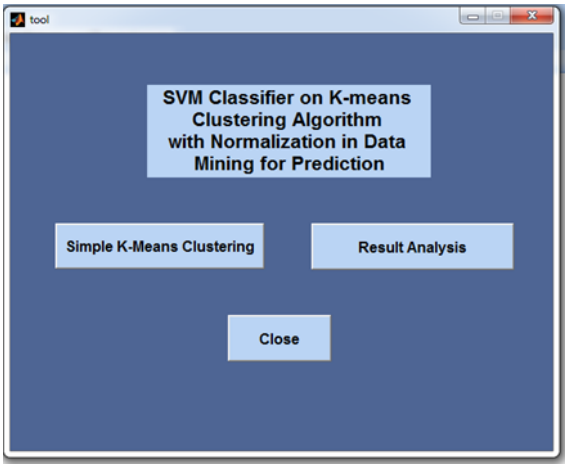

Figure 2: GUI of Simple K-Means

As shown in figure-1, This is the window for calculating the performance of existing algorithm and the K-means algorithm with the proposed modification. We are comparing the performance of algorithm on the basis of accuracy and execution time.

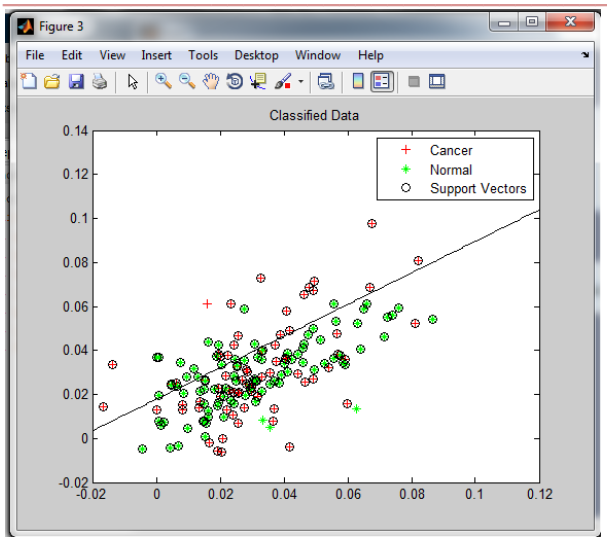

Figure 3 :Classified Data using k-mean algorithm

As shown in figure 8, The regions have been divided with the existing k-mean clustering algorithm. Here the dataset is divided in three regions. All these regions are different from each other as they all depict something different from the other one.And then SVM is applied to classify the dataset.

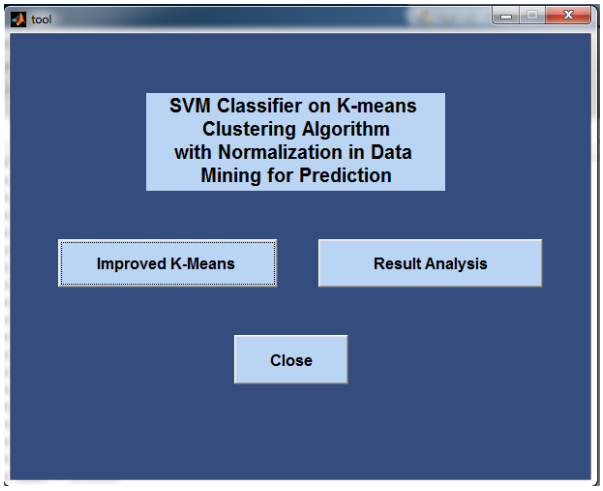

Figure 4: GUI Of Improved K-Means

As shown in figure 4, This is the window for calculating the performance of existing algorithm and the K-means algorithm with the proposed modification. We are comparing the performance of algorithm on the basis of accuracy and execution time.

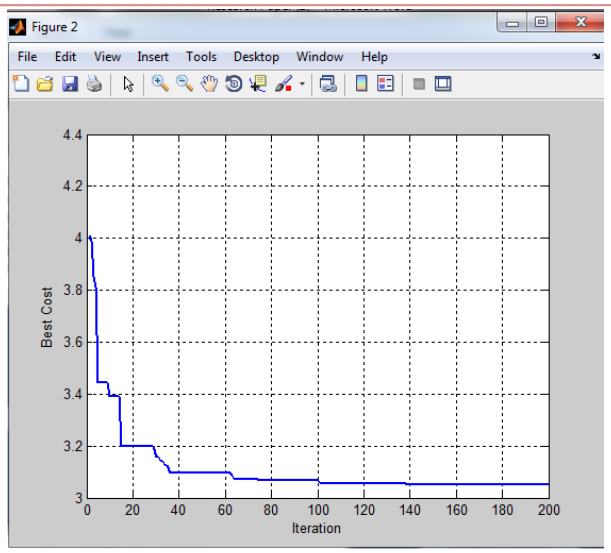

Figure 5: Best Cost and Iteration Analysis Graph

As shown in figure 5, A graph has been plotted on a 2D plane, for analysis of Cluster Quality Analysis. This graph has been plotted considering two factors, number of iterations that is on the X axis and best cost which is on the Y axis. After analyzing the graph, we can say that the quality of the third cluster s low.

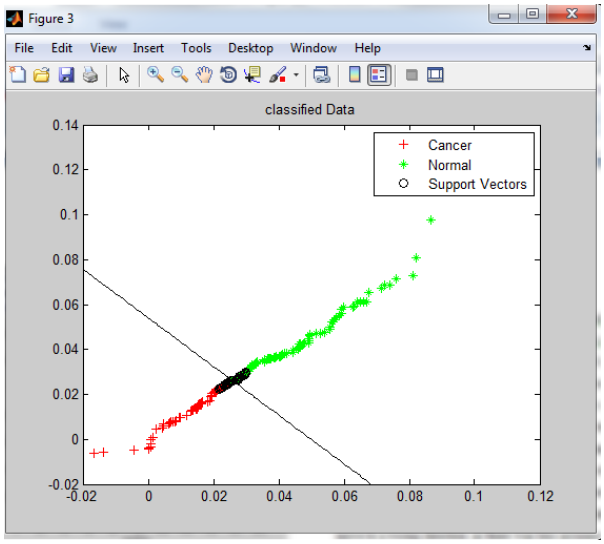

Figure 6 :Classified Data Using Improved K-Means

As shown in figure 6, The dataset has been divided into two separate regions. This graph plotted on a 2D plane, can be described easily as our technique has made it very easy to understand and predict after seeing the graph that the red dots that are below the line are indicting the cancerous and the green dots that above the line are non cancerous data.

A. RESULT

As it can be seen in the Table 1 shown below that there is a big variation between execution time of existing algorithm and proposed algorithm. And with this proposed algorithm the accuracy level gets almost double. So, it can be said that

the proposed modification in the existing k-means algorithm will give a huge improvement to the clustering techniques. The problem of accuracy level and execution time do not matter much, but when it comes to large dataset when there are millions of records, then it becomes enormously big problem. Because then the entire study of the dataset may move to a wrong direction as there was less accuracy level**.**

Table 1: Comparison with Existing Algorithm

Parameters	Simple K-Means Algorithm	Improved K-Means Algorithm
Execution Time	7.5089 e-02	8.1023 e-02
Accuracy	58.25%	91.64%

VII. CONCLUSION

In this paper, it is gathering that clustering is technique by which large number of datasets are divide in to small data collections with matching with similarity between them that converted data is called clusters the similar data which is collected in the form of cluster that are turned into information. The data clusters are different from each other as they are possessing some different values from each other. There are number of algorithms that work well for clustering the data that can divide a dataset into clusters. In this paper we have implemented technique for aimproved in K-Means Clustering Algorithm by changing some steps and adding some factors which are impacted on results for improvement. Here in this proposed modification, the K-Means clustering will vanish off the two major factor of K-Means clustering, that are accuracy level and algorithm run time consumed to clustering the dataset. Although when we use little amount of datasets those two factors consumption time and accuracy of result may not matter but when we use big amount of datasets that have impact on both result parameters i.e. accuracy and time, then little dispersion in accuracy level will matter a lot and can lead to a disastrous situation, if not handled properly. So, in last consideration the implemented idea of algorithm is more reliable and less time consuming with increased with improved of cluster quality.

VIII. FUTURE SCOPE

Following are the various possibilities which can be done in future

1. In future proposed technique can be compared with some other technique of classification technique
2. In future, technique will be proposed which can analyze various with other normalization technique instead of min max normalization.

IX. ACKNOWLDEMENT

I am extremely grateful and indebted to my parents and my colleague for being pillars of strength, for their unfailing moral support, and encouragement. I treasure their blessings and good wishes and dedicate this study to them.

I thank one and all who have been instrumental in helping me to complete this dissertation work.

REFERENCES

[1]. K. Rajalakshmi, D. S. Dhenakaran, and N. Roobin, "Comparative Analysis of K-Means Algorithm in Disease Prediction," International Journal of Science, Engineering and Technology Research (IJSETR), vol. 4, pp. 1-3, 2015.

[2]. O.Oyelade, O.Oladipupo, and I. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," arXiv preprint arXiv:1002.2425, 2010.

[3]. https://www.tutorialspoint.com/data_mining/

[4]. B.SundarV, T. Devi, and N. Saravanan, "Development of a Data Clustering Algorithm for Predicting Heart," International Journal of Computer Applications, vol. 48, pp. 8-13, 2012/06/30 2012.

[5]. https://in.mathworks.com/help/stats/kmeans.html

[6]. SumithaThankachan ,Suchithra, Data Mining & Warehousing Algorithms and its Application in Medical Science - A Survey, IJCSMC, Vol. 6, Issue. 3, March 2017, pg.160 – 168

[7]. Dr.B.Srinivasan , K.Pavya, A Study On Data Mining Prediction Techniques In Healthcare Sector, International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 03 | Mar-2016

[8]. Purvashi Mahajan, Abhishek Sharma, "Role Of K-meansAlgorithm in Disease Prediction", International Journal of Engineering And Computer Science, ISSN: 2319-7242, Vol.5, Issue 4, 2016, pp.16216-16217.

[9]. Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013

[10].https://en.wikipedia.org/wiki/K-means_clustering