Mine Blood Donors Information through Improved K Means Clustering

Paridhi Pachori
Department of Information Technology
College of Engineering, Bharati Vidyapeeth University, Pune

Abstract— The number of accidents and health diseases increasing at an alarming rate has resulted in a huge increase in the demand for blood. There is a necessity for the organized analysis of the blood donor database or blood banks repositories. Clustering analysis is one of the data mining applications and K-means clustering algorithm is the fundamental algorithm and traditional approach for modern clustering techniques. The K-means clustering is an iterative algorithm which attempts to find the distance from the centroid of each cluster to each and every data point at every iteration. This paper gives the improvement to the original k-means algorithm by improving the initial centroids with distribution of data. Results and discussions show that improved K-means algorithm produces accurate clusters in less computation time to find the donors information.

Key words: Alzheimer Disease, Mine Blood Donors Information, Clustering Algorithm

I. INTRODUCTION

Data Mining is defined as mining of knowledge from huge amount of data. Using Data mining we can predict the nature and behaviour of any kind of data. The past two decades has seen a dramatic increase in the amount of information being stored in the electronic format. This accumulation of data has taken place at an explosive rate.

Cluster Analysis of a data is an important task in Knowledge Discovery and Data Mining. Clustering is the process to group the data on the basis of similarities and dissimilarities among the data elements. Clustering is the process of finding the group of objects such that object in one group will be similar to one another and different from the objects in the other group. A good clustering method will produce high quality clusters with high intra cluster distance similarity and low inter cluster distance similarity. Similarity measure used is standard Euclidean distance but there can also be other distance measures such as Manhattan distance, Minkowski distance and many others.

The popular clustering approach can be partition based or hierarchy based, but both approaches have their own merits and demerits in terms of number of clusters, cluster size, separation between clusters , shape of clusters, etc. Some other approaches are also based on hybridization of different clustering techniques. Many Clustering algorithms use the center based cluster criterion. The center of a cluster is often a centroid, the average of all the points in a cluster.

This paper presents the partition based clustering method called as k- Means and its modified approaches with the experimental results. A partitioning method first creates an initial set of k partitions, where parameter k is the desired number of clusters as output. It then uses an iterative relocation technique that attempts to improve the partitioning of the data points. K-Means is a numerical, unsupervised iterative method. It is original and very fast, so in many practical applications this method has proved to be very effective way to produce good clustering results. But the computational complexity of original k- Means is very high, especially for large datasets. Moreover this algorithm results in different type of clusters depending upon the random choice of initial clusters.

II. LITERATURE SURVEY

In a web based information system for blood donation, Pavel Berkhin has performed extensive research in the field of data mining experiments and organized analysis of the blood bank repositories which is helpful to the health professionals for a better management of the blood bank facility. Arun K. Pujari et al. have worked to improve the performance of blood donation information analysis. In this paper, improved k-means clustering is adopted that improves the performance for determining the blood donors information based on the required blood group and location where needed.

Several attempts are made by the researchers to improve efficiency of the k-means clustering. The variants of the k-means clustering algorithm are K-modes and K-medoids. These algorithms replace the means with the modes and medoids. The K-modes algorithm handles the categorical data. These also give better performance based on the way we choose the initial modes and methods. For clustering the data, the k-means and the k-modes are integrated by the k-prototypes algorithm. Both the numeric and categorical attributes are taken into account to define the dissimilarity measure. Fahim A.M. and Yuan F worked for improving the performance of the k means by finding the initial centroids. The centroids obtained this by these methods provide consistent and accurate clusters for the given datasets. Abdul Nazeer K A brought forward an efficient method for assigning the data-points to the clusters. The original k-means algorithm computes the distances between the data points from all the centroids in every iteration, thus making this algorithm computationally very expensive. Another approach given by Fahim makes use of two distance functions for this purpose- one similar to the k-means algorithm and another one based on a heuristics to reduce the number of distance calculations. However as was the case in the original k-means algorithm, this algorithm too determines the initial centroids randomly, thus compromising the accuracy of the resultant clusters. Koteswara Rao proposes improved k-means algorithm using a $O(nlogn)$ heuristic method for finding the initial

centroids. In this method, the initial centroids are generated in accordance with the distribution of the data instead of generating them randomly.

III. K-MEANS CLUSTERING ALGORITHM

The K-means algorithm assigns each point to the cluster whose centroid (Centre) is in the nearest proximity (centroid is the average of all the points in the cluster). The centroid's coordinates are determined by calculating the arithmetic mean of all the points in the cluster separately for each dimension. The biggest asset of this algorithm is its speed, which allow for it to be run on large databases. Also the simplicity of this algorithm makes it all the more easy to use. But this algorithm has its own share of shortcomings associated with it. The inability of this algorithm to yield the same results with each run is a major disadvantage and this occurs as the derived clusters are not independent of the initial random assignments. Also, this algorithm considerably reduces the intra-cluster variance, but does not take measures to ensure the global minimum of the variance of the result.

There are two separate phases in the k-means algorithm- wherein the first phase involves the identification of k centroids, given that we know the number of clusters (K) beforehand thereby arriving at one centroid per cluster. Initially K points which are likely to be in different clusters have to be selected, which are then made the centroids of their respective clusters. The Euclidean distance is then calculated for each data point from each of the cluster centroid. Compare the values; find the closest centroid for the data point. Then bind the centroid and the data point which results in the completion of the first phase and then an early grouping is done. At this stage, the new centroids have to be determined by calculating the mean value for each cluster. After we get k new centroids, then a new binding is to be created between the same data points previously used and the nearest new centroid, thus generating a loop. Because of this loop, the k centroids are prone to a change in their positions in a step by step manner. This process is repeated until convergence criteria is meat means the centroids of clusters are do not move anymore.

A. Algorithm: The K-Means Clustering Algorithm

1) Input:

D = {d1, d2,.....,dn} //set of *n* data items.

k// Number of desired clusters

2) Output:

A set of *k* clusters.

3) Steps:

- Arbitrarily choose *k* data-items from D as initial centroids;
- Repeat
- Assign each data item *di* to the cluster which has the closest centroid;
- Calculate the new mean of each cluster.

IV. APPLICATIONS

1) Often used as an exploratory data analysis tool
2) In one –dimension, a good way to quantize real –valued variables into k non-uniform buckets.
3) Used on acoustic data in speech understanding to convert waveforms into one of k categories (known as Vector Quantization)
4) Also used for choosing color palettes on old fashioned graphical display device

V. ADVANTAGES AND DISADVANTAGES

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments (the k-means++ algorithm addresses this problem by seeking to choose better starting clusters). It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement for the concept of a mean to be definable which the case is not always. For such datasets the k-medoids variants is appropriate. An alternative, using a different criterion for which points are best assigned to which centre is k-medians clustering.

VI. PROPOSED SYSTEM

The proposed system finds the blood donors information and satisfies the need by these steps:

1) Collect data from the blood bank. Apply the improved K-means Algorithm to classify the number of blood donors through the blood group and location where it needed.
2) Extract the Mail ids from the resultant cluster which satisfies the criteria of blood group and location.
3) Forward the message to the donors.

The user interface for the project would look like as shown in figures 1.

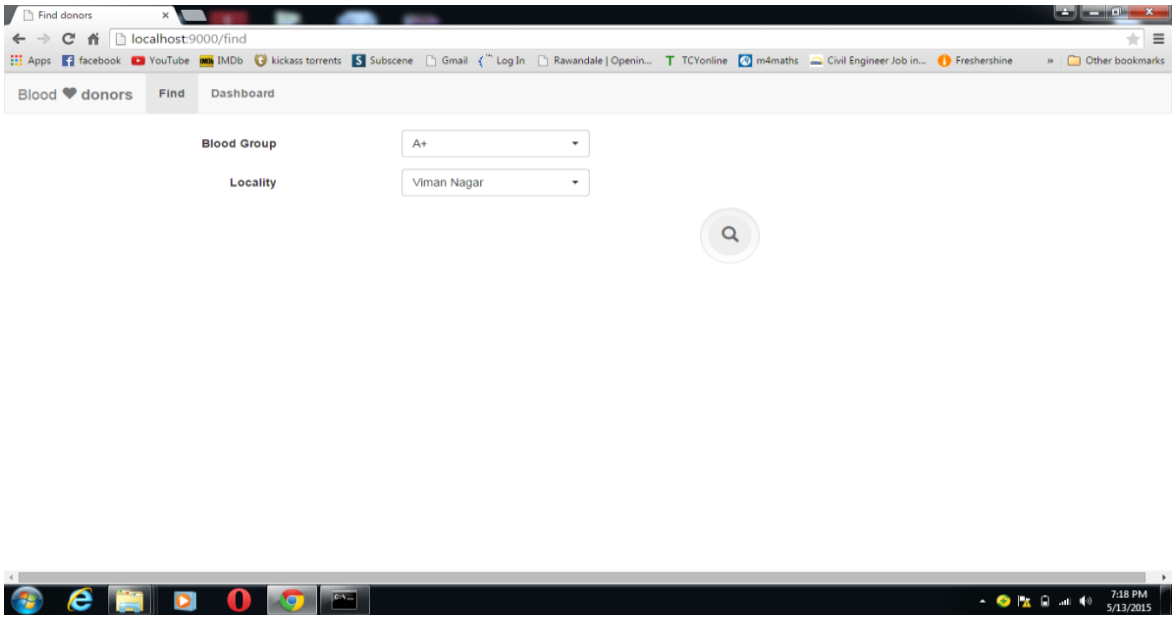

Fig. 1: shows the dashboard where the donor enters all the data and information, based on which the clusters are made.

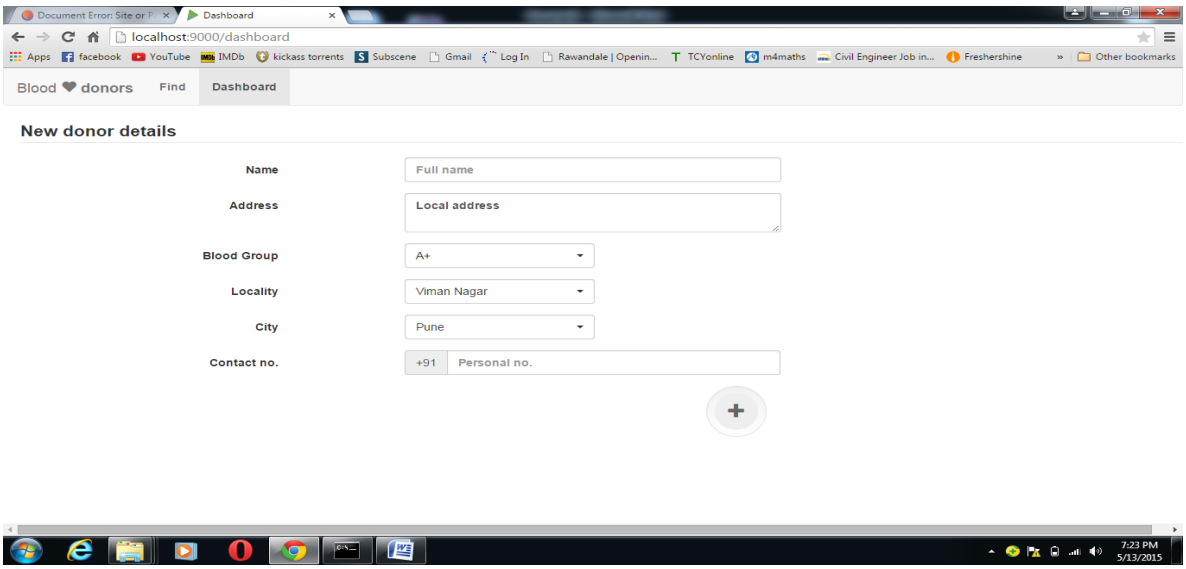

Fig. 2: shows the "find" functionality. When the donor gives an entry in Blood Group and Locality field, relevant output is fetched from the database.

In this methodology, the improved clustering algorithm, deals with the multi-dimensional data values. Each data point di may contain multiple attributes such as di1, di2…dim, where m is the number of attributes or columns in each data value. In such cases we first determine the column with maximum range [7], where range is the difference between the maximum and the minimum element in the column.

Then determine the initial centroids [10] as range of the Column is divided by the number of clusters and sum with the minimum value of the column. i.e., assume we have the two dimensional dataset, then the initial centroid are

Cx=MAXx-MINxK+1+clusterid*MINx Cy=MAXy-MINyK+1+clusterid*MINy

Hence the data points are divided into k-equal partitions. For each iteration, we calculate the Euclidean distance between the data point and each centroid.

For all geometrical problems, the standard distance metric used is the Euclidean distance, which is just the direct distance between any two points and can be easily measured with the help of just a ruler, be it in two-dimensional space or three-dimensional space. It is a default distance measure [3] for the k-means clustering algorithm. Clustering problems and clustering text widely use of Euclidean distance, which is a true metric as it, satisfies all the above stated four conditions. Also the Euclidean metric is the default distance metric which is used in the k-means algorithm. The distance measured by the Euclidean metric is also known by 'as-the-crow-flies' distance.

This distance from a point X (X1, X2, etc.) to a point Y (Y1, Y2, etc.) is the square root of the sum of the squares of the differences between the corresponding values of the two data points is necessary for finding the Euclidean distance. The new centroid is found by calculating the mean value of distance of the all data points of that cluster. Each data point di is assigned to the cluster having the closest centroid. The distance between the data points and the centroid is measured by using the Euclidean distance.

Improved K-means Clustering Algorithm is outlined as below:

A. Algorithm 2: K-means Clustering Algorithm with improved initial centroids:

1) Input:

D = {d1, d2... dn} // set of *n* data items.
K // Number of desired clusters.

2) Output:

A set of *k* clusters.

3) Steps:

- Calculate the initial centroids according to the formula given above and set the cluster with that centroid. Repeat
- Initially assign the each data point to the cluster
- Update the centroid value by calculating the mean of that cluster until all data points are assigned to any one of the clusters. Repeat.
- Assign each data item *di* to the cluster which has the closest centroid;
- Calculate new mean of each cluster; until convergence criterion is met.

This algorithm gives the better performance and accurate results within less time compared to the literature [1]

VII. RESULT ANALYSIS

Now in order to tabulate the Speed test, we performed the test in two ways. In the first test we consider the results when the K-Means Clustering Algorithm is used, and in the other where no clusters are formed. We have considered two code snippets. In the first code snippet (Fig 3), clusters are not formed and the data is fetched from the database directly without using any algorithm. The time recorded for fetching the records is - 586ms

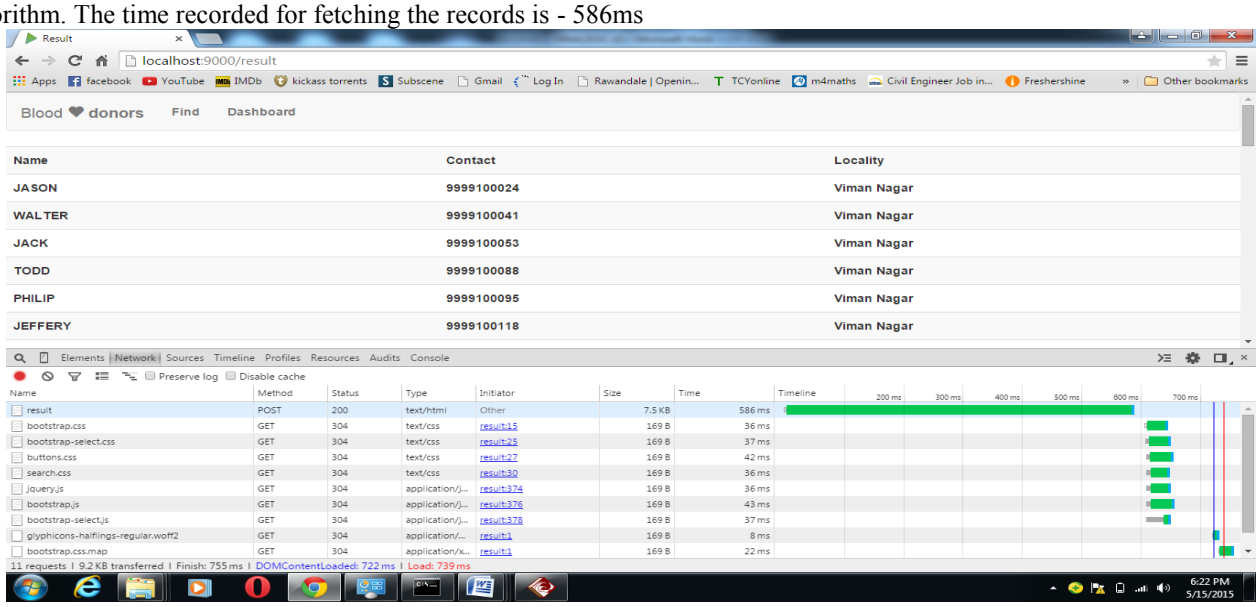

Fig. 3: Results

In the second code snippet (Fig 4) clusters are formed as the K means Clustering Algorithm is used. The data is fetched from the server's cache (where the clusters are formed) and the time taken to fetch is recorded as- 77 ms .

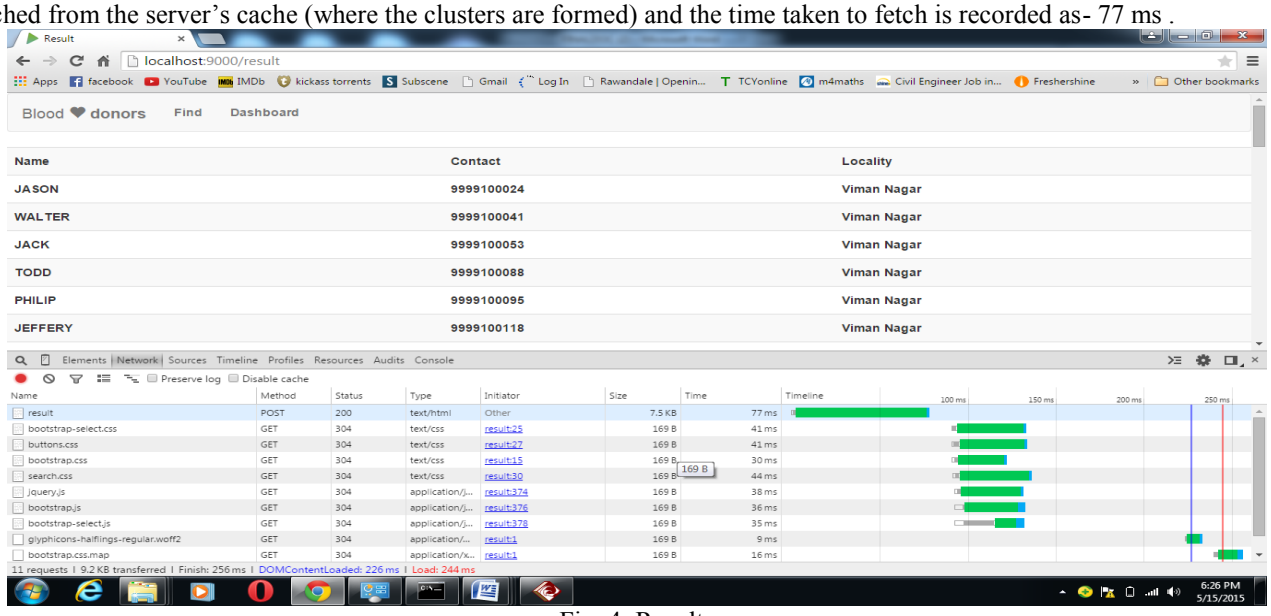

Fig. 4: Results

The tests are carried on the same database , with same parameters , and the results show that K means clustering algorithm displays the result in 77ms whereas with no clustering time taken is - 586ms and hence there is a difference of- 509 ms.

The test carried for 2 other parameters depicted results with similar patterns. In all three cases the K means algorithm is able to fetch the result in less time compared to the non -clustering methodology.

VIII. CONCLUSION

This K-means algorithm has many real time applications, but its performance cannot be guaranteed as it takes the initial centroids randomly. Also the computational complexity of the original k-means algorithm is alarmingly high considering the need to reassign the data points every time the loop runs. In this method, we are taking the initial centroids in a meaning full way and assign data points to the clusters in a distributed manner. This results in better accuracy compared to the classic K-means Algorithm. Results show that the improved initial centroids with k-means clustering algorithm give the accurate and efficient results. A limitation of the improved k-means algorithm is that it still takes the input as number of clusters and takes the location in a static way.

IX. FUTURE SCOPE

In this paper, blood donors data retrieval system has been constructed which provides fast retrieval of information by using K – means clustering algorithm. It can be used in crisis situations to provide accurate, easy and fast retrieval of the data from a database. Hence it can be used in organizations of various levels for easy and fast retrieval of data from varied datasets. The clustering algorithm splits the data into a fixed number of clusters based on the parameters and stores it into the cache which results in speedy retrieval of data as compared with the relational database. This improved algorithm can be effectively used by various organizations like in hospitals, colleges, residential societies, and even governmental organizations like the municipalities. All the members who enter their correct information shall be able to get information about the potential blood donors in the nearest possible location in emergency.

REFERENCES

[1] Jaipur National University, Jaipur, “Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool”
[2] N. Koteswara Rao, G. Sridhar Reddy “ Discovery of Preliminary Centroids Using Improved K- Means Clustering Algorithm”
[3] K. RAJENDRA PRASAD and Dr. P.GOVINDA RAJULU, “A survey on clustering technique for datasets using efficient graph structures” International Journal of Engineering Science and Technology Vol. 2 (7), 2010, 2707-2714.
[4] Jiawei Han M. KAMBER, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier.
[5] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.
[6] McQueen J, “Some methods for classification and analysis of multivariate observations,” Proc. 5th Berkeley Symp. Math. Statist. Prob., (1):281–297, 1967.
[7] Pang-Ning Tan, Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007, Transactions on Information Theory, 28(2): 129-136.
[8] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, “An Efficient enhanced k-means clustering algorithm,” Journal of Zhejiang University, 10(7):1626–1633, 2006.
[9] Huang Z, “Extensions to the k-means algorithm for clustering large data sets with categorical values,” Data Mining and Knowledge Discovery, (2):283–304, 1998.
[10] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, “A New Algorithm to Get the Initial Centroids,” Proc. of the MINE BLOOD DONORS INFORMATION THROUGH IMPROVED KMEANS CLUSTERING 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.
[11] Wen-ChanLee and Bor-Wen Cheng; ‘An Intelligent system for improving performance of blood donation’, Journal of Quality, Vol. 18, Issue No. II, 2011.
[12] Arun. K. Pujari (2001): Data Mining Techniques, Universities Press
[13] Abdul Nazeer K A, Sebastian M P, “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm,“ Proceedings of the International Conference on Data Mining and Knowledge Engineering, London, UK, 2009.
[14] Nidhi Singh and Divakar Singh, “Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4119-4121.