A NOVEL METHOD TO AVOID UNNECESSARY COMPUTATION AND EFFICIENT PRUNING OF CLUSTERS USING K-MEDIAN PARTITIONING TECHNIQUE FOR HIGH DIMENSIONAL BIOLOGICAL DATASET

Bhavya T

Department of Computer Science and Engineering,

East West Institute of Technology, Visvesvaraya Technological University, (India)

ABSTRACT

K-median clustering is mostly used to gain more knowledge in life science dataset and to obtain tighter clusters with less outliers. The two widely used distance measurements Pearson correlation distance and standardized Euclidean distance are used as they both yield same k-median clustering for similar sets of k initial centroids. Proposed method avoids unnecessary calculations and obtains tighter clusters. Triangular inequality is used to define norms property and to measure distance.

Keywords: K-Median, Mining Methods And Algorithm, Distance Measurements

I. INTRODUCTION

Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster. Clustering is the process of making group of abstract objects into classes of similar objects. The main advantage of Clustering over classification is that, it is adaptable to changes and help single out useful features that distinguished different groups. Clustering is unsupervised grouping of objects into classes without any a prior understanding of the dataset to be analyzed. K-median clustering is mostly used for processing huge datasets as they have less computational complexities when compared with hierarchical clustering. It is a variation of k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median.

Traditional K-median clustering is briefly explained as below:

1) For each point, find the closest cluster center. Initial k centroids are found using many initialization methods as proposed by [1], [2]. The initialization method proposed by Bradley and Fayyad [3] is mainly used here as report said report by Celebi et al. [1].
2) Compute the new set of cluster centers computing the median of the cluster. In other words, for each dimension compute the median value for that dimension over all points in the cluster.
3) Terminate if the stopping condition is met else repeat step 2.

Euclidean distance is the distance between two points in Euclidean space Euclidean distance is sensitive to scaling while standardized Euclidean distance and Pearson correlation distance are insensitive to scaling and they produce same clustering result for similar sets of k initial centers as standardized Euclidean distance is the square root of Pearson correlation distance[4], [5]. Euclidean distance has to be normalized because variables measured in large units will dominate small valued units which contribute less. Triangular inequality is the sum of lengths of any two sides must be greater than or equal to the length of other side. Pearson correlation coefficient violates triangular inequality.

Pearson correlation coefficient is concentrated more in this paper to develop a robust method to avoid unnecessary computation and efficient pruning is possible as Euclidean distance is used.

II. LITERATURE SURVEY

Many papers have been published on accelerating k-means algorithm, in different research communities. Some of these papers are briefly described here.

A heuristic method was proposed to accelerate k-means clustering by avoiding unnecessary computational complexities [7] using Pearson Correlation distance. Outperformed pruning using Euclidean distance. Later comparison of computational performance was done against five best algorithms such as Lloyd's algorithm [6], BoostKCP (boundA) [7], BoostKCP (boundB) [7], Elkan's algorithm [8], Hamerley's algorithm [9], [10]. Used initialization method proposed by Bradley and Fayyad [3].

The problem of clustering a set of points to minimize maximum intercluster distance is studied. An approximation algorithm called O(kn), where n is the number of points and k is the number of clusters, which assures that an objective function value is two times the optimal solution value [2]. This is one of the efficient initialization method to calculate better k centroids.

Accelerated algorithm was developed to avoid unnecessary distance calculations by applying triangular inequality. Triangle inequality was applied in two ways, by tracking down lower and upper bounds between data points and centers in order to calculate distance. The upper bounds u(x) and lower bounds l(x,c) are tight for most points x and centers c at the starting of each iteration and as the resulting centers for next iteration might have slightly changed the number of distance calculation is reduced. Number of distance computation is nke where n is the number of data points, k is the number of clusters and e is the number of iterations required [8]

One of the efficient k-means clustering algorithm to process Euclidean distance by using triangular inequality was proposed by Greg Hamerly. The algorithm uses updated distance bounds and triangle inequality to avoid individual distance computations and to skip the loop which iterates over k centers. This is achieved by maintaining two distance bounds (upper and lower bounds) per data point for its two closest center. Main advantage of this algorithm is that it uses small memory when compared with other accelerated algorithms. [9].

J. Drake and G. Hamerly proposed algorithm similar to Elkan's and Hamerly's algorithm but took a variable number of lower bounds, which is adjusted at runtime. Adaptive tuning mechanism has been built which help to proceed without knowing optimal b lower points in advance [10].

Bradley's and Fayyad's initialization method starts by randomly partitioning the data set into J subsets. These subsets are clustered using k-means initialized by MacQueen's second method. MacQueen's second method chooses the centers randomly from the data points so that they pick points from dense regions, points good to be centers. Produces J sets of intermediate centers each with K points, later combined into superset then clustered by k-means J times, each time initialized with a different center set. Final centers are from these centers set which gives the least SSE (Sum of Squares for Error) [3].

III. METHOD

Proposed method leverage the complementary strengths of Kazuki Ichikawa's and Shinichi Morishita's heuristic method by using K-median clustering, which automatically forms tighter cluster and less outliers are detected. Pearson correlation distance is mainly concentrated to avoid unnecessary computational complexities and to outperform pruning of clusters standardized Euclidean distance is used.

The analysis of raw data involves three main steps:

1) Preprocessing: Normalization of data and calculating pairewise similarity values between elements.

Normalization in clustering is a preprocessing step to prevent attributes with large ranges from dominating the distance calculations and avoid numerical instabilities during computations. Two commonly used normalization schemes are linear scaling to unit range (min-max normalization) and linear scaling to unit variance (z-score normalization.

2) Clustering.

3) Evaluation of the result.

Definition of Pearson's correlation coefficient: To measure the distance between two d dimensional vectors x=(x[1],…,x[d]), y=(y[1],…,y[d]),

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{\boldsymbol{x}[i]-\overline{\boldsymbol{x}}}{\sigma_{\boldsymbol{x}}}\right)\left(\frac{\boldsymbol{y}[i]-\overline{\boldsymbol{y}}}{\sigma_{y}}\right),$$

where $\overline{\boldsymbol{x}}$ denotes the average of x[1],…,x[d], and $\sigma_{\boldsymbol{x}}$ is standard deviation.

Pearson's correlation coefficient ranges between -1 and 1i.e., -1 ≤ p(x,y) ≤ 1. High correlation ranges from 0.5 to 1.0 or -0.5 to 1, medium correlation ranges from 0.3 to 0.5 or -0.3 to 0.5 and low correlation ranges from 0.1 to 0.3 or -0.1 to -0.3. The Pearson correlation distance, dis(x,y) is defined as 1 – p(x,y). Pearson correlation distance is 0 when both x and y are dissimilar. Range of distance is 0 ≤ dis(x,y) ≤ 2.

Definition of standardized Euclidean distance: Let dis_SE(x,y) denote

$$\sqrt{\sum_{i=1}^{d}\left(\frac{\mathrm{x}[i]-\overline{\boldsymbol{x}}}{\sigma_x}-\frac{\boldsymbol{y}[i]-\overline{\boldsymbol{y}}}{\sigma_y}\right)^2},$$

the standardized Euclidean distance between two d dimensional vectors x and y.

The square root of the Pearson correlation is proportional the standardized Euclidean distance.

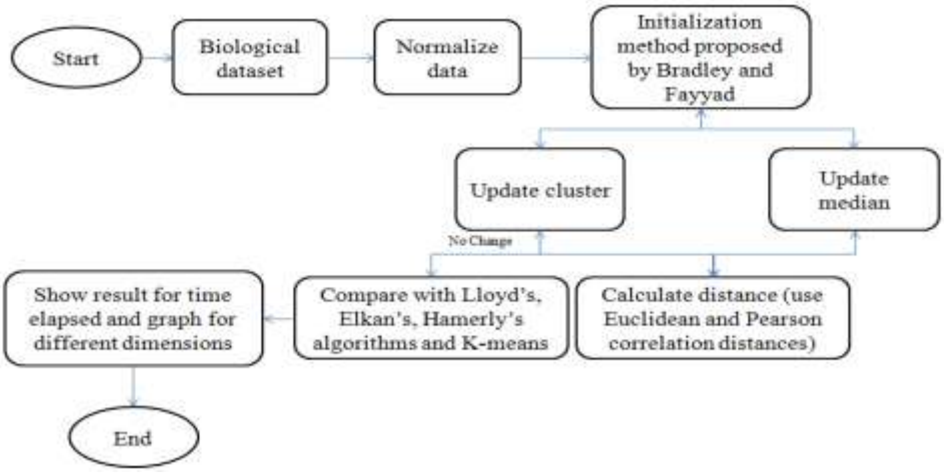
Start
Biological dataset
Normalize data
Initialization method proposed by Bradley and Fayyad
Update cluster
Update median
No Change
Show result for time elapsed and graph for different dimensions
Compare with Lloyd's, Elkan's, Hamerly's algorithms and K-means
Calculate distance (use Euclidean and Pearson correlation distances)
End

Figure 1: Architecture of Proposed Method

Fig. 1 describes the architecture of the proposed method, K-median clustering is used along with Pearson correlation and standardized Euclidean distance later on compared with Lloyd's, Elkan's and Hamerly's algorithms.

IV. CONCLUSION

The overall objective is to study the k-median clustering technique and compare with Lloyd's, Elkan's, Hamerly's algorithms and K-means clustering technique to devise a novel method using high dimensional biological datasets resulting in tighter clusters with less outliers.

REFERENCE

[1] H. A. Kingravi, M. E. Celebi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," Expert Syst. Appl., vol. 40, pp. 200–120, 2012.

[2] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," Theoretical Comput. Sci., vol. 38, pp. 293–306, 1985.

[3] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering," in Proc. 15th Int. Conf. Mach. Learn., 1998, pp. 91–99.

[4] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," IEEE Trans. Knowl. and Data Eng., vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[5] K. L. Clarkson, "Nearest-neighbor searching and metric space dimensions," Nearest-Neighbor Methods Learning Vis.: Theory Practice, pp. 15–59, 2006.

[6] S. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[7] K. Ichikawa and S. Morishita, "A Simple but Powerful Heuristic Method for Accelerating k-Means Clustering of Large-Scale Data in Life Science" IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 11, No. 4, July/August 2014

[8] C. Elkan, "Using the triangle inequality to accelerate k-means," in Proc. Int. Conf. Mach. Learn., 2003, p. 147-153

[9] G. Hamerly, "Making k-means even faster," in Proc. Symp. Data Mining, 2010, pp. 130–140.

[10] J. Drake and G. Hamerly, "Accelerated k-means with adaptive distance bounds," in Proc. 5th NIPS Workshop Optimization Mach.Learn., 2012.