PERFORMANCE ASSESSMENT OF IMPROVED FARTHEST FIRST CLUSTERING ALGORITHM ON SMARTPHONE SENSORS DATA

[1]Dr.M.Jayakameswaraiah,[2]Dr.K.Suresh Kumar Reddy,[3]Dr.S.Ramakrishna

[1]Assistant Professor,[2]Academic Consultant,[3]Professor

[1]School of Computer Science and Applications, Reva University, Bangalore, Karnataka, India.

[2]Department of Computer Science and Engineering, S.V.U College of Engineering, Sri Venkateswara University,Tirupati,Andhra Pradesh,India.

[3]Department of Computer Science, Sri Venkateswara University, Tirupati,Andhra Pradesh, India.

---

*Abstract:*Research on effective methods to deal with ever larger data sets has been gaining importance in recent years. The goal of clustering is to organize a data collection into clusters, such that items within each cluster are more similar to each other than to items in other clusters. While supervised clustering assumes that some information is available concerning the membership of data items to predefined classes, unsupervised clustering does not require a priori knowledge of data contents. There are many applications of unsupervised clustering in computer vision, pattern recognition, information retrieval, data mining, etc. The objective of this research work is focused on the ethical cluster creation of smartphone sensors data and analyzed the performance of partition based algorithms. Dataset consist of all the information gathered during the network connection established with wristwatch and smartphone, which needs to be, analyzed the performance of the proposed improved farthest first clustering algorithm.

***IndexTerms*-Clustering algorithms, Improved Farthest First algorithm, Density Based Cluster, Filtered Cluster, K-Means, Data Mining.**

I. INTRODUCTION

Data Mining discovers unknown relationships in data; in fact it is part of a wider process called "knowledge discovery". Knowledge discovery defines the phases which should be completed to ensure getting significant results through research. The objective of DM process is to obtain information out of a dataset and convert it into a comprehensible outline. Also, it includes the following: data reprocessing, data management, database aspects, visualization and complexity considerations, online updating, inference and model considerations, interestingness metrics. On the other hand, the actual data mining assignment is the semi-automatic or automatic exploration of huge quantities of information to extract patterns that are interesting and previously unknown. Clustering is one of the most common untested data mining methods that explore the hidden structures embedded in a dataset. Clustering is the method of making group of objects into classes of similar objects[6,10]. A cluster of data objects can be treated as one group. While doing the cluster analysis, first partition the set of data into groups based on data similarity and then allocates the label to the groups[4].

II. LITERATURE REVIEW

Forperforming comparison analysis dataset is obtained from UCI data repository. These repositories are very helpful in data analysis. This data can directly process in data mining tools for predicting the results.

2.1 Clustering Techniques

Clustering has wide applications. It is often used as an individual data mining tool to observe the characteristics of each cluster and to focus on a particular set of clusters for further analysis. Clustering not only can act as an individual tool, but also can serve as a preprocessing step for other algorithms which would then operate on the detected clusters. The clusters can be formed based upon various parameters depending upon the clustering method [9,11]. Different Clustering methods have a database of n objects or data tuples.

- Partitioning methods
- Hierarchical based methods
- Density based method
- Grid-based methods
- Model-based methods
- Clustering high-dimensional data
- Constraint-based clustering

2.1.1 Filter Based Clustering

This category of clustering method is for filtering theinformation or any pattern which are essentially needed. The filtration is carried out based on the keywords that aresupplied or some relevant information. Jiang-She Zhang et al proposed a clustering algorithm for the processingof the images. They are computationally stable andinsensitive to initialization. They also produce consistent clusters. Thomas et al proposed a collaborativefiltering which is a combination of the correlation andsingular value decomposition (SVD) to improve accuracy [7,13].

2.1.2 Density Based Clustering

The density based clustering (DBC) groups the objects mainly based on the density of the objects that arereachable and connective. Li Tu et al proposed aframework called D-stream for clustering using the density based approach. Mitra et al suggested anonparametric data reduction scheme. The procedurefollowed here is separating the dense area objects fromless dense area with the aid of an arbitrary object[1,3,15]. Thedensity based clusters (DBC) are robust to noise but thedatasets are problematic and requires high denselyconnected data.

2.1.3 Centroid Based clustering

K-Means is a centroid based clustering method. Itpartitions the dataset into various clusters based on themean distance. It is one of the simplest forms ofunsupervised algorithm. The main objective of thisalgorithm is to reduce the squared error. Tapas et al identified that the algorithm works faster as the separationbetween the cluster increases. This algorithm is applicablefor the segmentation of images and data compression.Kanungo et al proposed that the K-means algorithmruns faster as the separation between the cluster increases [2,5,14].

III. IMPLEMENTATION OF IMPROVED FARTHEST FIRST ALGORITHM

Farthest first algorithm proposed by Hochbaum and Shmoys 1985 has same procedure as k-means, this also chooses centroids and assign the objects in cluster but with max distance and initial seeds are value which is at largest distance to the mean of values, here cluster task is different, at initial cluster we get link with extraordinary Session Count, like at cluster-0 more than in cluster-1, and so on.Working as labeled here, it also describes initial seeds and then on basis of "k" number of cluster whichwe need to know prior[3,15]. In farthest first it takes point *Pi* then chooses next point *P*1 which is at maximum distance. $P_i$is centroid and p_1, p_2, …,$p_n$are points or objects of dataset belongs to cluster from equation 1.

$$\min\{\max dist(p_i, p_1), \max dist(p_i, p_2)\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Farthest first actually solves problem of k-Centre and it is very efficient for large set of data. In farthest first algorithm we are not finding mean for calculating centroid, it takes centroid arbitrary and distance of one centroid.For each $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$ in D that is labeled by *m* categorical attributes, we use $f(x_{ij}|D)$ to denote the frequency count of attribute value $x_{i,j}$in the dataset. Then, a scoring task is designed for evaluating each point, which is defined as:

$$Score(X_i) = \sum_{j=1}^{m} f(x_i|D)$$

Step-1: Farthest first traversal (D: data set, *k*: integer)
Step-2: Randomly select first center
Step-3: //select centers
Step-4: For (i= 2,…,*k*)
Step-5: For (each remaining point) {calculate distance to the current center set;
Step-6: Select the point with maximum distance as new center
Step-7: Apply Euclidean Distance function on each cluster
Step-8: //assign remaining points
Step-9: for (each remaining point)
Step-10: Calculate the distance to each cluster center using Manhattan distance formula
Step-11: put it to the cluster with minimum distance
Step-12: repeat the steps until each cluster remains

3.1 Euclidean Distance

This Euclidean distance between the two documents d_i, d_j can be intended as

$$Euclidean\ Distance(d_i, d_j) = \sqrt{\sum_{k=1}^{n} (d_{ik} - d_{jk})^2}$$

Where, n is the number of positions present in the vector space model.Euclidian distance gives the dissimilarity between the two documents. If the distance is lesser it indicates they are more similar else dissimilar.

3.2 Manhattan Distance

The Manhattan distance between the two documents d_i, d_j can be intended as

$$Manhattan\ Distance(d_i, d_j) = \sum_{k=1}^{n} |d_{ik} - d_{jk}|$$

3.3 Accuracy Measure

To determine the accuracy of the clusters created by clustering algorithms, F-measure is used. Every cluster created by clustering algorithms is considered as the result of a query and the documents in these clusters are treated as set of retrieved documents. The documents in each category are considered as set of appropriate documents. The documents in the cluster that are relevant to that cluster are set of relevant documents retrieved. Precision, Recall and F-measureare calculated for every cluster[12,16,17]. Thecomplete accuracy of the clustering algorithm is the average ofthe accuracy of all the clusters. Precision, recall and F-measure are considered as follows:

Precision = relevant documents retrieved /retrieved documents

Recall = relevant documents retrieved / relevant documents

F-Measure = (2 * Precision * Recall)/ (Precision + Recall)

IV. Results And Discussions

In this section, by smartphone sensors dataset a series of experiments are carried out to categorize the documents into predefined number of categories by using improved farthest first algorithm. The initial centroids for k-means algorithm are chosen randomly by using farthest neighbors. The accuracy of the clusters and proficiency of the algorithm is examined. In this work the tests are carried out with one of smartphone sensors dataset collected from UCI and KDD repositories. Thesmartphone_sens consists of 18354 Instances and 13 attributes that are timestamp, AccelerationX, AccelerationY,AccelerationZ, MagneticFieldX,MagneticFieldY,MagneticFieldZ, Z-AxisAgle(Azimuth), X-AxisAngle(Pitch), Y-AxisAngle(Roll), GyroX, GyroY,GyroZ.

Table 4.1: Performance comparison of clustering algorithms

Algorithm	Number of Clusters	Number of Clustered Instances	Percentage of Clustered Instances
K-Means	Cluster 0	10644	58%
	Cluster 1	7710	42%
Density Based Clusterer	Cluster 0	10775	59%
	Cluster 1	7579	41
Filtered Clusterer	Cluster 0	10644	58%
	Cluster 1	7710	42%
Improved Farthest First Clusterer	Cluster 0	17999	98%
	Cluster 1	355	2%

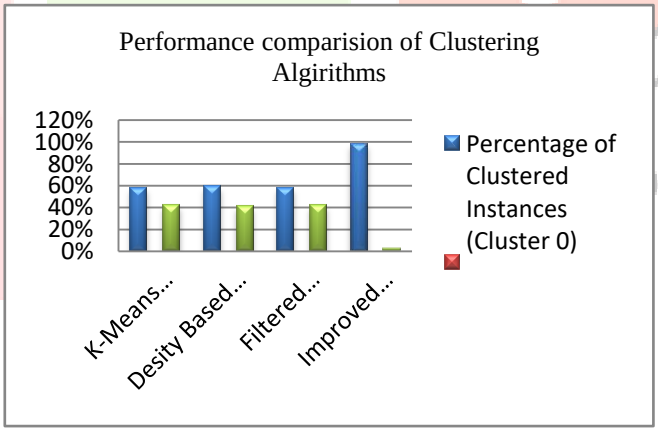

Figure 4.1: Performance comparison of Clustering Algorithms on Cluster 0

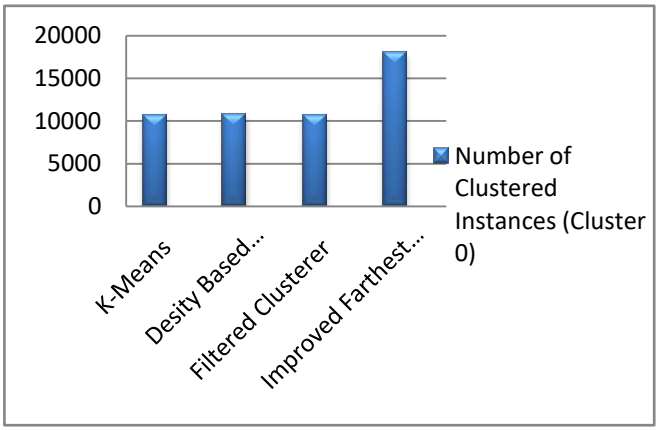

Figure 4.2: Number of Clustered Instances on Cluster 0

V. CONCLUSTION

Clustering has wide applications. It is often used as an individual data mining tool to observe the characteristics of each cluster and to focus on a particular set of clusters for further analysis. Clustering not only can act as an individual tool, but also can serve as a preprocessing step for other algorithms which would then operate on the detected clusters. In this experimental research we estimated the performance of the clustering algorithms namely K-Means, Density Based clusterer, Filtered clusterer and our proposed Improved Farthest First Clusterer are evaluated on smartphone sensor data which is taken from the UCI-Machine learning repository. In this research we applied various clustering algorithms on the smartphone sensors data set and compared the clustering accuracy performances using data mining tool. But it is observed that our proposed algorithm gives 98% of clustering accuracy on cluster 0. That means it contributes better results than other standard clusters in data mining system.

REFERENCES

[1]. DanasinghAsir Antony Gnana Singh, Subramanian Appavu Alias Balamurugan, Epiphany JebamalarLeavline, "Improving the Accuracy of the Supervised Learners using Unsupervised based Variable Selection", Asian Journal of Information Technology, Volume 13,issue 9,ISSN: 530-537,2014.

[2]. D. Isa, V. P. Kallimani, and L. H. Lee, "Using the self-organizing map for clustering of text documents", Expert Systems With Applications, vol. 36, pp. 9584-9591, 2009.

[3]. D. Ryabko and J. Mary. A binary-classification-based metric between time-series distributions and its use in statistical and learning problems. Journal of Machine Learning Research, volume 14, ISSN: 2837-2856, 2013.

[4]. D.JanakiSathya, "Development of Intelligent System Based on Artificial Swarm Bee Colony Clustering Algorithm for Efficient MassExtraction from Breast DCE-MR Images", Int. Journal. on Recent Trends in Engineering and Technology, Vol. 6, No. 1, 2011.

[5]. Guha, Meyerson, A. Mishra, N. Motwani, and O. C, "Clustering data streams: Theory and practice", IEEE Transactions on Knowledge and Data Engineering, vol. 15, pp. 515-528, 2003.

[6]. Joshua Zhexue Huang, Michael K. Ng, HongqiangRong, and Zichen Li, "Automated Variable Weighting in k-Means Type Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 27, NO. 5, PP. 657-668, 2005.

[7]. K. Mumtaz1 and Dr. K. Duraiswamy, "A Novel Density based improved k-means Clustering Algorithm – Dbk-means", International Journal on Computer Science and Engineering, Vol. 02, No. 02,ISSN: 0975-3397 213, 2010.

[8]. K Mohamed, E Côme, L Oukhellou, "Clustering Smart Card Data for Urban Mobility Analysis", IEEE Transactions on Intelligent Transportation Systems, Volume: 18, Issue: 3, 2017.

[9]. M.Jayakameswaraiah, S.Ramakrishna, "A Study on Prediction Performance of Some Data Mining Algorithms", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014, ISSN: 2321-7782.

[10]. M.Jayakameswaraiah, Prof.S.Ramakrishna, "Development of Data Mining System to Analyze Cars using TkNN Clustering Algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 7, July 2014,ISSN: 2278 – 1323.

[11]. Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C, " Application of k- means Clustering algorithm for prediction of Students Academic Performance", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.

[12]. R. Xu and D. Wunsch, "Survey of Clustering Algorithms", IEEE Transactions on Neural networks, vol. 16, no. 3, 2005.

[13]. Shi Na, Liu Xumin, "Research on k-means Clustering Algorithm", IEEE Third International Conference on Intelligent Information Technology and Security Informatics, 2010.

[14]. T.Sridevi, "An Innovative Algorithm for Feature Selection Based on Rough Set with Fuzzy C-Means Clustering", Journal of Theoretical and Applied Information Technology, Vol. 68, No.3, 2014.

[15]. Vadeyar, Deepshree A., and H. K. Yogish. "Farthest First Clustering in Links Reorganization", International Journal of Web & Semantic Technology Volume 5, issue 3, 2014.

[16]. Yang Yang, Zhigang Ma, Yi Yang, FeipingNie, Heng Tao Shen, "Multitask Spectral Clustering by Exploring Inter task Correlation", IEEE Transactions on Cybernetics, Volume: 45, Issue: 5, 2015.

[17]. Zhong W, Altun G, Harrison R, Tai PC, Pan Y. "Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property", IEEE Transactions on Nano Bioscience, Volume 4, issue 3, ISSN:255–65, 2005.