IMPROVEMENT IN K-MEANS CLUSTERING ALGORITHM FOR DATA CLUSTERING

Prof.Dr.k.Rajeswari*, Omkar Acharya†, Mayur Sharma‡, Mahesh Kopnar§and Kiran Karandikar ¶
*Pimpri Chinchwad College of Engineering, Nigdi, Pune 411044
Email: raji.pccoe@gmail.com
†BE Computer Engg, Pimpri Chinchwad College of Engineering, Nigdi, Pune 411044
Email: omkaracharya5757@gmail.com
‡BE Computer Engg, Pimpri Chinchwad College of Engineering, Nigdi, Pune 411044
Email:mayursharma60@yahoo.com
§BE Computer Engg, Pimpri Chinchwad College of Engineering, Nigdi, Pune 411044
Email: mkopnar@gmail.com
‡BE Computer Engg, Pimpri Chinchwad College of Engineering, Nigdi, Pune 411044
Email:khkarandikar@gmail.com

***Abstract*—The set of objects having same characteristics are organized in groups and clusters of these objects are formed known as Data Clustering.It is an unsupervised learning technique for classification of data. K-means algorithm is widely used and famous algorithm for analysis of clusters.In this algorithm, n number of data points are divided into k clusters based on some similarity measurement criterion. K-Means Algorithm has fast speed and thus is used commonly clustering algorithm. Vector quantization,cluster analysis,feature learning are some of the application of K-Means.However results generated using this algorithm are mainly dependant on choosing initial cluster centroids.The main shortcome of this algorithm is to provide appropriate number of clusters.Provision of number of clusters before applying the algorithm is highly impractical and requires deep knowledge of clustering field. In this project, we are going to propose an algorithm for improvement in the initializing the centroids for K-Means algorithm. We are going to work on numerical data sets along with the categorical datasets with the n dimensions. For similarity measurement we are going to consider the manhattan distance ,Dice distance and cosine distance. The result of this proposed algorithm will be compared with the original K-Means.Also the quality and complexity of the proposed algorithm will be checked with the existing algorithm**

Index Terms- Data Clustering, K-Means,unsupervised learning,centroid.

I. INTRODUCTION

Data analysis underlies many computing applications, either as part of their online operations or in a design phase. Data analysis procedures can be divided as either confirmatory or exploratory, based on the availability of appropriate models for the data source, but a primary element in both types of procedures (whether for hypothesis formation or decisionmaking) is the grouping, or the classification of measurements based on either (i) goodness-of-fit to a postulated model, or (ii) natural groupings (clustering) revealed through the analysis. Cluster analysis is the organization of a collection of patterns (usually represented as a point in a multidimensional space, or a vector of measurements) into clusters based on similarity. Clustering technique is a significant unsupervised method for grouping of data. Clustering is a method in which we make the clusters of items that are somehow related in same features. The aim of the clustering is to provide a grouping of similar records of data. Clustering is often confused with classification, but there is some dissimilarity between the two. A very simple measure is the intra-cluster distance, which, as in the K-means algorithm, needs to be minimized for better clustering results. The term clustering is used in several research communities to describe methods for grouping of untagged data or unlabeled data. The communities have different terminologies and assumptions for the components of the clustering process and the context in which clustering is in use. The production of a truly comprehensive survey would be a monumental task given the sheer mass of literature in the area of clustering. The accessibility of the survey might also be questionable given the need to reconcile very different vocabularies and assumptions regarding clustering in the various communities. Typical pattern clustering activity involves the following steps [Jain and Dubes 1988]::

1) Data Point Representation.
2) Data Points Similarity Measurement.

978-1-4799-6892-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCUBEA.2015.205

3) Clustering or grouping.
4) Data abstraction (if needed).
5) Assessment of output (if needed).

II. Traditional k means

In this algorithm, n number of data points are divided into k clusters based on some similarity measurement criterion. K-Means Algorithm has fast speed and thus is used commonly clustering algorithm. Vector quantization,cluster analysis,feature learning are some of the application of K-Means. **Steps**

1) Pick a number (K) of cluster centers
2) Assign every data point (e.g., gene) to its nearest cluster center
3) Move each cluster center to the mean of its assigned data points (e.g., genes)
4) Repeat 2-3 until convergence

Calculate distance of data points from centroids using:

$$D(Xp,Cj)=\sqrt{\sum_{i=1}^{d}(Xpi - Cji)^2}$$

Where Xp denotes the pth data vector, j denotes the centroid vector of cluster j, and d subscripts the number of features of each centroid vector.

Recalculate new centroid from generated clusters using:

$$Cj = \frac{1}{Nj}\sum \forall xp \in cj\, Xp$$

Where Nj is the total number of data vectors in cluster j and j is the subset of vectors that form cluster Cj.

A. Time Complexity

The problem of finding the global optimum is NP-Hard.
Time complexity is O(n*k*i)
Where n=Total no of elements.
k=No of cluster iteration.
i=iterations.

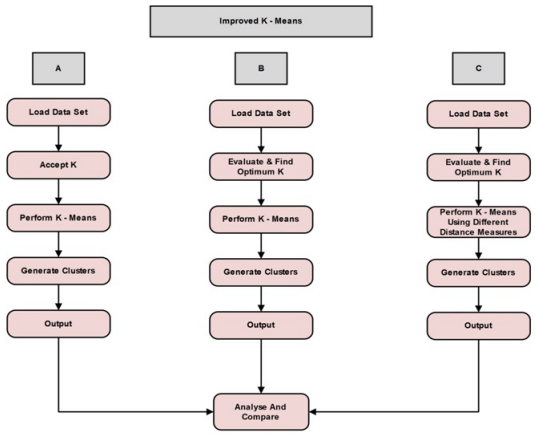
III. Proposed k-means

The main shortcome of K-means algorithm is to provide appropriate number of clusters.Provision of number of clusters before applying the algorithm is highly impractical and requires deep knowledge of clustering field. In this project, we are going to propose an algorithm for improvement in the initializing the centroids for K-Means algorithm.

A. Mathematical Model:

Let S be the system,
S = I, Fn, C, S, F
where,
I = Set of input M, Fe
M = term-document matrix n*m
n = number of datapoints.
m = number of attributes.
Fe=set of initial centroid c1, c2, , ck
k = number of centroids.
Fn = f(D1), f(D2), f(D3)
D1 = is to check whether the datapoint present in Cluster C1
D2 = is to check whether the datapoint present in Cluster C2
D3 = is to check whether the datapoint present in Cluster C3
where,
C1, C2, and C3 are the number of clusters.
D1, D2 and D3 are the distances calculated from each datapoint to the initial centroid.
Constraints C = Data should be numeric.
Success S = Dataset is divided into defined number of cluster.
Failure F = Error Rate i.e. the original dataset and the clusters obtained after clustering give much difference the intracluster distance is very high.

B. Flow Diagram



C. Algorithm

1) Select two centroids from the dataset i.e lowest centroid point and the highest centroid point.
2) After choosing the centroids we create two clusters with members which are dissimilar to

each other.

Input: T:The set of n Data points with attributes T1,T2,Tn

where, n=no of attributes. All attributes are numeric.

Output: Suitable No of Clusters with n Data points distributed properly.

Steps:

a) Compute sum of attributes values of each Data points (to find the points in the datasets which are farthest apart).
b) Take Data points with lowest and highest values of the sum as the initial centroids .
c) Create initial partitions (clusters) using Cosine , Sorensen-Dice and Manhattan Distance measurement formulas between every Data points and the initial centroids.
d) Find distance of every Data points from the centroid in both the initial partitions .Take l=lowest of all distances.
e) Find the new centroids for the partitions created in step c.
f) Compute the distance of every Data points from the new cluster centers and find the outliers depending on the following objective functions:
g) If the distance of the Data points from the cluster means ¡ l then not a outlier.
h) Compute the new centroids for the clusters.
i) Calculate the distance of every outlier from the new cluster centroids and find the outliers which are not satisfying in step f.
j) Let O=OL1,OL2,.,OLp be the set of the outliers obtained in step h(the value of K depends on numbers of outliers).
k) Repeat until (A==NULL)

- Create a new cluster for the set B ,by taking mean value of its members as centroid.
- Find the outliers of this clusters ,depending on the objective function in step 6.
- If no of outliers =n then Create a new cluster with one of the outliers as its members and test every other outlier for the objective function in step f.
- Find the outliers if any.
- Calculate the distance of every outlier from the centroid of the existing clusters and adjust the outliers in the existing which satisfy the objective in step f.
- A=O1,O2.Oq be the new set of outliers.(value of q depends upon the no of outliers.)

l) If newly generated cluster centroid are much closer then adjust the value of l manually unless you get more seperated clusters.

IV. Conclusion

For applying traditional k means algorithm,we need to provide number of clusters initially.This requires deep knowledge about mining field and dataset. And if the input is given incorrectly the result will be affected and it may not be accurate . Thus a system is suggested using Improved K-means algorithm to remove the drawbacks of existing K-Means, i.e. provision of number of clusters before application. We will compare the results of the existing K-Means algorithm with the Proposed K-Means algorithm along with the quality of clusters and complexity of algorithm.

V. References

1) Yi-Tung Kao, Erwie Zahara and I-Wei Kao, A Hybridized Approach to Data Clustering, Proceedings of the 7th Asia Pacific Industrial Engineering and Management Systems Conference 2006 17-20 December 2006, Bangkok, Thailand.
2) Tapas Kanungo,David Mount,Nathan Netanua,Christian Piatko,Ruth Silverman and Angela Wu,An Efficient K-Means Clustering Algorithm; Analysis and Implementation IEEE transactions on Pattern Analysis and Machine Intelligence,Volume 24,No 7,July 2002.
3) M.V.B.T.Santhi,V.R.N.S.S.V.SaiLeela,P.U.Anitha, D.Nagamalleswari, Enhancing K-Means Clustering Algorithm,International Journal on Computer Science And Technology(IJCST) Vol 2,Issue 4,Oct-Dec 2011
4) Chunfei Zhang ,Zhiyi Fang, An Improved K-means Clustering Algorithm , Journal of Information And Computational Science 10: 1 (2013) 193199,January 2013
5) Shi na,Guan Yong,Liu Xumin, Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm ,Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on 2-4 April 2010.
6) Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques.
7) Jiangang Qiao,Yonggang Lu,A new Algorithm for choosing initial Cluster Center for K-means,International Conference on Computer Science And Electronic Engineering (ICCSEE 2013).