Comparing and Selecting Appropriate Measuring Parameters for K-means Clustering Technique

Shreya Jain, Samta Gajbhiye

Abstract— ***Clustering is a powerful technique for large scale topic discovery from text. It involves two phases: first, feature extraction maps each document or record to a point in a high dimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters. Hence to improve the efficiency & accuracy of mining task on high dimensional data the data must be pre-processed by an efficient dimensionality reduction method. Recently cluster analysis is popularly used data analysis method in number of areas. K-Means is one of the well known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids. In this paper, a certain k-means algorithm for clustering the data sets is used and the algorithm outputs k disjoint clusters each with a concept vector that is the centroid of the cluster normalized to have unit Euclidean norm. Also in this paper, we deal with the analysis of different sets of k-values for better performance of the k-means clustering algorithm.***

Keywords: Data Mining, Text Mining, Clustering, K-Means Clustering, Silhouette plot .

I. INTRODUCTION

Data mining (the analysis step of the **knowledge discovery in databases** process, or KDD), a relatively young and interdisciplinary field of computer science is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The goal of data mining is to extract knowledge from a data set in a human-understandable structure and involves database and data management, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structure, visualization and online updating.

Text mining also called intelligent text analysis, text data mining, or knowledge discovery in text — uncovers previously invisible patterns in existing resources. To perform analysis, decision-making, and knowledge management tasks, information systems use an increasing amount of unstructured information in the form of text. This data influx, in turn, has spawned a need to improve the text mining technologies required for information retrieval, filtering, and classification. People who are involved in doing research can systematically analyze multiple research papers, e-books and other documents, and then swiftly determine what they contain. For example in an HR department, a CV which matches a particular job specification from amongst a million CV in a database may not be the simplest of tasks. [2]

Cluster analysis is one of the major data analysis methods widely used for many practical applications in emerging areas. **Clustering** is the process of finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns.

Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-median in which the objective is to minimize the sum of distances to the nearest centre and the geometric k-centre problem in which the objective is to minimize the maximum distance from every point to its closest centre [4].

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by their centroids, by minimizing the square error function. Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen.

Section 1 of the paper deals with the introductory concepts of Data Mining , Text Mining , Clustering and K-Means Clustering. Section 2 describes the Clustering technique and K-Means clustering algorithm. Section 3 describes the problems related in selection of wrong measuring parameters in K-Means clustering. Section 4 describes the experimental activities and corresponding result discussions which is followed by conclusions in Section 5

Revised Manuscript Received on March 2012.

Shreya Jain, M.E. Scholar, Computer Science & Engg Department, Shri Shankaracharya Technical Campus, Bhilai, India, Mobile No. -99074 19538 (e-mail: shreyajain.0312@gamil.com).

Samta Gajbhiye, Sr. Associate Professor, Computer Science & Engg Department, Shri Shankaracharya Technical Campus, Bhilai, India, Mobile No. -98261 05305 (e-mail: samta.gajbhiye@gamil.com).

II. RELATED WORK

Several attempts were made by researchers to improve the effectiveness and efficiency of the K-means algorithm. Yuan *et al*. (2004) proposed a systematic method for finding the initial centroids. However, Yuan's method does not suggest any improvement to the time complexity of the K-means algorithm. Belal *et al*. (2005) proposed a new method for cluster initialization based on finding a set of medians extracted from a dimension with maximum variance. Zoubi *et al*. (2008) proposed a new strategy to accelerate K-means clustering by avoiding unnecessary distance calculations through the partial distance logic. Fahim *et al*. (2009) proposed a method to select a good initial solution by partitioning dataset into blocks and applying K-means to each block. Here the time complexity is slightly more. . Though the above algorithms can help finding good initial centers for some extent, they are quite complex and some use the K-means algorithm as part of their algorithms, which still need to use the random method for cluster center initialization.

A. Clustering

Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

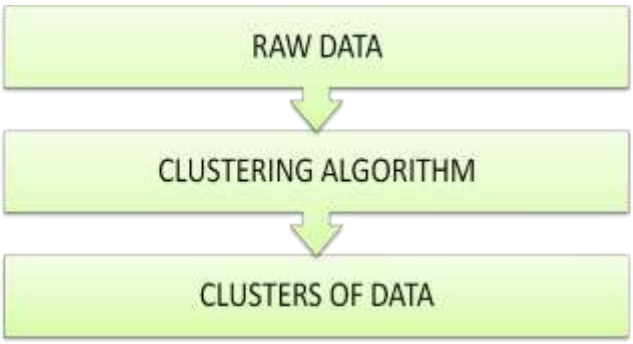

Fig 1 Clustering

CLUSTERING problems arise in many different applications, such as data mining and knowledge discovery, data compression and vector quantization, and pattern recognition and pattern classification [4].

B. K-Means Clustering

The K-means algorithm is one of the partitioning based, nonhierarchical clustering methods. Given a set of numeric objects ***X*** and an integer number *k*, the K-means algorithm searches for a partition of ***X*** into *k* clusters that minimizes the within groups sum of squared errors. The K-means algorithm starts by initializing the *k* cluster centers. The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster. The processes of re-assigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers.

In paper 7 , the steps of the K-means algorithm are :
1. Initialization: choose randomly *K* input vectors (data points) to initialize the clusters.
2. Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.
3. Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster.
4. Stopping rule: repeat steps 2 and 3 until no more change in the value of the means [7]

K-Means Algorithm

Step 1 : Enter the number of clusters and number of iterations, which are the required and basic inputs of the K-means clustering algorithm.

Step 2: Compute the initial centroids by using the Range Method shown in equations below:
c_i = ((max X - min X) / k)* n
c_j = ((max Y - min Y) / k)* n
The initial centroid is C(c_i, c_j).
Where max X, max Y, min X and min Y represent maximum and minimum values of X and Y attributes respectively.
K represents the number of clusters
i, j and n vary from 1 to k where k is an integer.

In this way, we can calculate the initial centroids; this will be the starting point of the algorithm. The value (maxX – minX) will provide the range of X attribute, similarly the value (maxY – minY) will give the range of Y attribute. If both the attributes have zero value then this formula will not work. The value of k must be at least 1 if k is zero then again it will give an error, the division by zero. The value of n varies from 1 to k. The number of iterations should be small otherwise the time and space complexity will be very high and the value of initial centroids will also become very high and may be out of the range in the given dataset.

Step 3: Calculate the distance by Euclidean's distances. On the basis of these distances, generate the partition by assigning each sample to the closest cluster.
Euclidean Distance Formula:
$d (x_i , x_j) = \sum (x_{ik} - x_{jk})$
Where d(xi, xj) is the distance between xi and xj. xi and xj are the attributes of a given object, where i and j vary from 1 to N where N is total number of attributes of a given object. i,j and N are integers.

Step 4: Compute new cluster centers as centroids of the clusters, again compute the distances and generate the partition. Repeat this until the cluster memberships stabilizes .

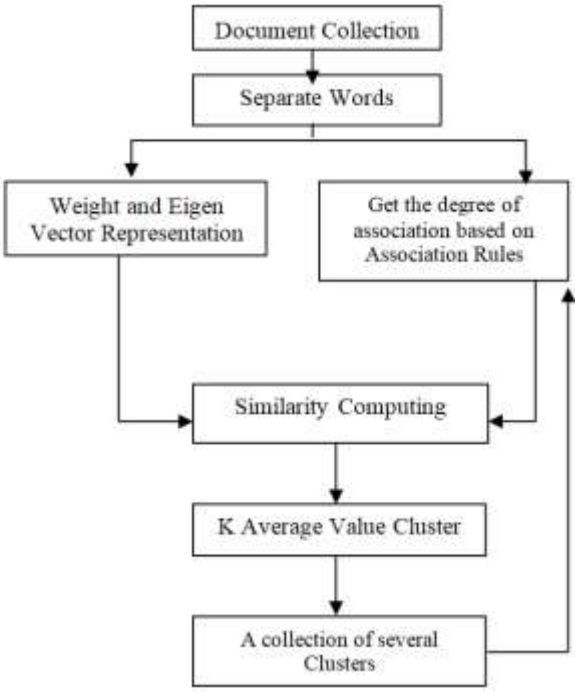
Document Collection
Separate Words
Weight and Eigen Vector Representation
Get the degree of association based on Association Rules
Similarity Computing
K Average Value Cluster
A collection of several Clusters

Fig 2. Flow Diagram of K- Means Clustering

K-Means clustering is preferred due to several reasons which are as follows :

- With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
- K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.
- Fast, robust & easier to understand.

III. PROBLEMS IN K-MEANS CLUSTERING

Selection of value of K

The performance of a clustering algorithm may be affected by the chosen value of K. Therefore, instead of using a single predefined K, a set of values might be adopted. It is important for the number of values considered to be reasonably large, to reflect the specific characteristics of the data sets. At the same time, the selected values have to be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering.

Selection of centroid values

When the K-means algorithm is applied to data with a uniform distribution and K is increased by 1, the clusters are likely to change and, in the new positions, the partitions will again be approximately equal in size and their distortions similar to one another. The evaluations carried out in reference showed that, when a new cluster is inserted into a cluster (K ¼ 1) with a hypercuboid shape and a uniform distribution, the decrease in the sum of distortions is proportional to the original sum of distortions. This conclusion was found to be correct for clustering results obtained with relatively small values of K. In such cases, the sum of distortions after the increase in the number of clusters could be estimated from the current value.

IV. EXPERIMENTAL ACTIVITIES AND RESULT DISCUSSIONS

The sample dataset which we are using for testing our results is Iris dataset and in this dataset 150 examples of data are taken with 3 different labels and 6 different attributes according to which clustering for different values of k is done.

Selecting Value of k with silhouette plot

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster.

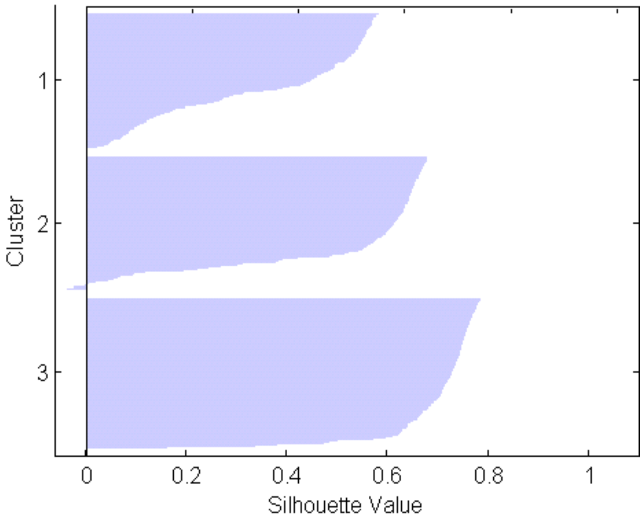

Fig 3: Silhouette Plot with k= 3

From Fig 3, we can see that most points in the third cluster have a large silhouette value, greater than 0.6, indicating that the cluster is somewhat separated from neighbouring clusters. However, the first cluster contains many points with low silhouette values, and the second contains a few points with negative values, indicating that those two clusters are not well separated.

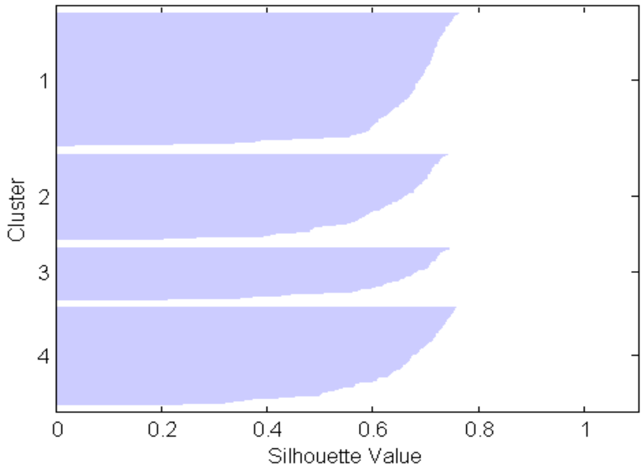

Fig 4: Silhouette Plot with k= 4

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

In Fig 4 , the value of k is increased to see whether it results in finding better clusters. So here we get the Silhouette value for all the clusters above 0.6 so with this value of k, it results in better formation of clusters.

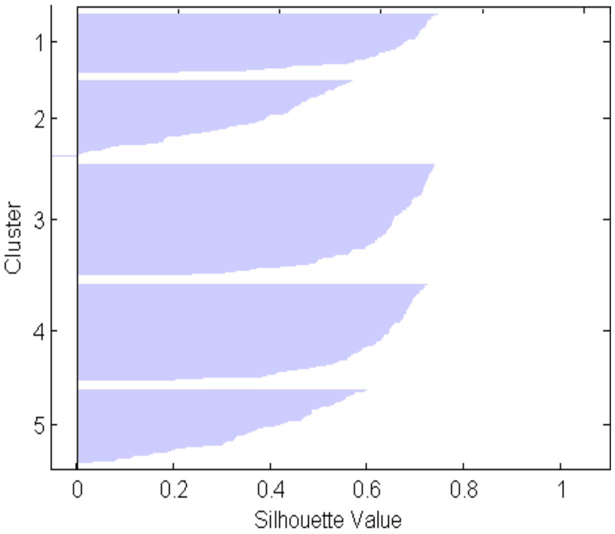

Fig 5: Silhouette Plot with k= 5

In Fig 5 , we get the Silhouette value for 2 clusters less than 0.6 so this shows that this value of k is not suitable for clustering .

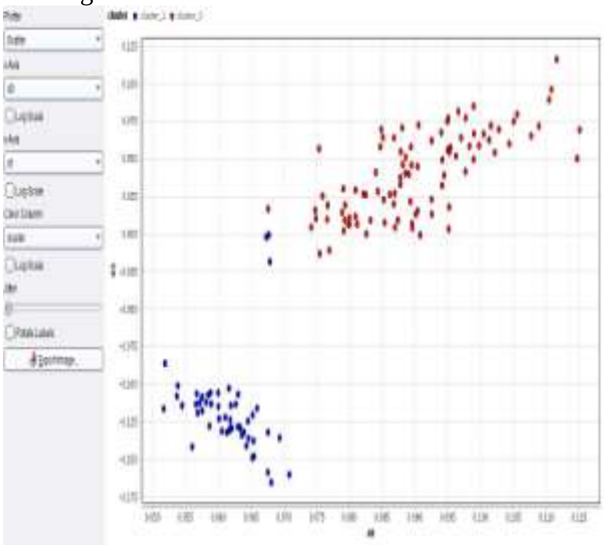

Fig 6: Clustering with k=2

Fig 6 shows the clustering of data when k=2.

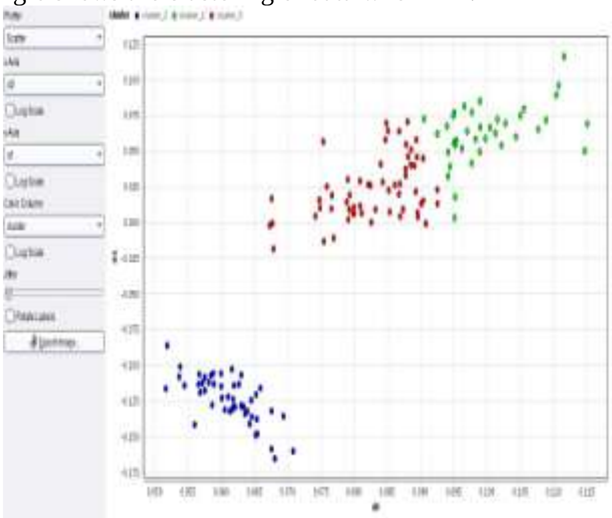

Fig 7 Clustering with k=3

Fig 7 shows the clustering of data when k=3.

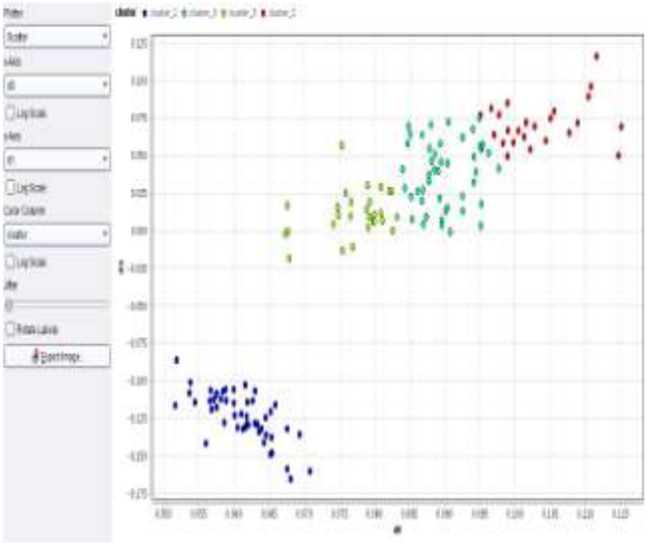

Fig 8 Clustering with k = 4

Fig 8 shows the clustering of data when k=4.

V. CONCLUSIONS

In this paper, different techniques like Silhouette plot etc. has been used which helps in finding the better value of k and also the randomly chosen initial centroid values. K-Means can be extended to ensure that every cluster contains at least a given number of points. Our experimental results shown that by observing the Silhouette values for different values of k, we can choose the better value of k and initial centroid values so that efficient and good clusters of data can be obtained. Again the method to find the initial centroids may not be reliable for very large dataset. Methods for refining the computation of initial centroids are worth investigating.

ACKNOWLEDGMENT

I am thankful to Prof. Kamal Mehta(Head of Department), Computer Science & Engineering for giving thoughtful suggestions during my work. I owe the greatest debt and special respectful thanks to Dr. P. B. Deshmukh (Director), Shri Shankaracharya Technical Campus, Bhilai for their inspiration and constant encouragement that enabled me to present my work in this form.

REFERENCES

[1] Vishal Gupta & Gurpreet S.Lehal ,"*A Survey of Text Mining Techniques &Application* " ,Journal of Emerging Technologies in Web Intelligence ,Aug 2009.

[2] Atika Mustafa, Ali Akbar, and Ahmer Sultan , " *Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization",* International Journal of Multimedia and Ubiquitous Engineering , Vol. 4, No. 2, April, 2009 .

[3] Raymond J.Mooney & Razvan Bunescu , "*Mining Knowledge from Text using Information Extraction ".*

[4] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu , "*An Efficient k-Means Clustering Algorithm: Analysis and Implementation" , IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 24, NO. 7, JULY 2002.

[5] Bjornar Larsen and Chinatsu Aone *," Fast and Effective Text Mining Using Linear-time Document Clustering",SRA International,* 2000.

[6] ZHEXUE HUANG , "*Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values* " **,***Kluwer Academic Publishers* ,1998.

[7] Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath, Milu Acharya ," *A hybridized K-means clustering approach for high dimensional dataset*" , International Journal of Engineering, Science and Technology, Vol. 66 2, No. 2, 2010, pp. 59-66.

[8] Charles Elkan , " *Using the Triangle Inequality to accelerate k- Means*" , Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003 .

AUTHOR PROFILE

Shreya Jain has completed her B.E. in Computer Science & Engineering from Bhilai Institute of Technology,Durg with 82.4% under Chhattisgarh Swami Vivekanand Technical University. Now she is pursuing her M.E in Computer Technology & Applications at Shri Shankaracharya Technical Campus, Bhilai.

Samta Gajbhiye is Sr. Associate Professor in Department of Computer Science & Engg. , Shri Shankaracharya Technical Campus, Bhilai under Chhattisgarh Swami Vivekanand Technical University, Bhilai .She received her Masters degree in Computer technology from NIT, Raipur. Her research areas include Data Mining, Cryptography etc.