Proceedings of
National Conference on Innovative Trends in Computer Science Engineering (ITCSE-2015)
held at BRCMCET , Bahal on 4th April 2015

A review paper on use of different weighing measures in K-Means Clustering for Interclass and Intraclass distance

[1]Pallavi, [2]Harish Bedi, [3]Isha Nagpal
[1,3]PPIMT, Hisar , Haryana
[2]BRCM,Bahal Bhiwani, Harynana
[1]pallavihsr.bedi@gmail.com , [2]hbedi@brcm.edu.in , [3]isha.parnami@gmail.com

Abstract— The objective of our research will be to compare different k-means algorithm distance measures. K-Means clustering makes the distances of objects in the same cluster as small as possible, but another objective that the distances of objects from different clusters is not taken into account. This paper will provide a review of comparative analysis how these two objectives will be satisfied. Different cost function for weighting measure for k-means clustering algorithm will be compared.

Keywords—Clustering, K-Means algorithm.

I.INTRODUCTION

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another and very dissimilar to object in other clusters. Dissimilarity is based on the attributes values describing the objects. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.

Firstly the set of data is portioned into groups based on data similarity (e g Using clustering) and the then assign labels to the relatively small number of groups. Several clustering techniques are there: partitioning methods, hierarchical methods, density based methods, grid based methods, model based methods, methods for high dimensional data and constraint based clustering. Clustering is also called data segmentation because clustering partitions large data sets into groups according to their similarity.

Clustering can be used for outlier detection where outliers may be more interesting then common cases e g Credit card fraud detection, monitoring of criminal activities in electronic commerce. Clustering is a pre-processing step for other algorithms such as characterization, attribute subset selection and classification, which would then operate on the detected clusters and the selected attributes or features.

II.K-MEANS CLUSTERING

The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the three steps below until convergence. Iterate until stable (= no object move group):

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance (find the closest centroid). This is showed in figure 1.1 in steps.

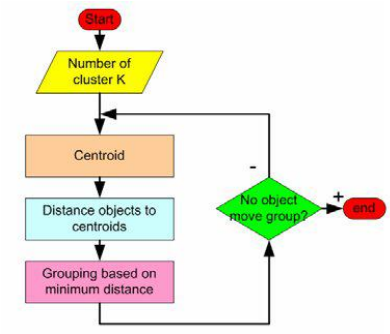

FIGURE 1: K-MEANS CLUSTERING PROCESS

III.TYPICAL REQUIREMENTS OF CLUSTERING

The main requirements of clustering in data mining are Scalability, ability to deal with different types of attributes, Discovery of clusters with different shapes, Minimal requirement for domain knowledge to determine input parameters, Ability to deal with noisy data, Incremental Clustering and insensitivity to the order of input records, High dimensionality, Constraint based Clustering and Interpretability and usability.

IV.PREVIOUS WORK AND ANALYSIS

In traditional clustering algorithm, the objective of clustering, which is to make the distance of objects from different clusters
as large as possible, is not taken into account and the dataset with mixed data can't be classified efficiently. They propose an Improved Weight Entropy algorithm by modifying the cost function of the entropy weighting k-means clustering algorithm in their research. In our research we will compare different algorithms having different distance measures.

The second objective of clustering, which is to make the distance of objects from different clusters is as large as possible, is not taken into account in traditional clustering algorithm. They propose an improved algorithm bymodifying the cost function of the entropy weighting k-means clustering algorithm in this paper. The proposed algorithm adjusts the existing cost function by adding a variable relevant the distances between the mean of all objects and the mean of each cluster.

V.PROPOSED WORK

Clustering can be used for outlier detection where outliers may be more interesting then common cases e g Credit card fraud detection, monitoring of criminal activities in electronic commerce. Clustering is a pre-processing step for

other algorithms such as characterization, attribute subset selection and classification, which would then operate on the detected clusters and the selected attributes or features. Contributing area of research include data mining, statistics, machine learning, special database technology, biology and marketing. Clustering is an unsupervised learning. Unlike classification, it does not rely on predefined classes and class labels training examples. Hence we say clustering is learning by observation rather than learning by examples.

In this review our objective will be to comparison of different k-means algorithm distance measures using WEKA Tool. K-Means clustering makes the distances of objects in the same cluster as small as possible, but another objective that the distances of objects from different clusters is not taken into account. This paper will provide a review of comparative analysis how these two objectives will be satisfied. Different cost function for weighting measure for k-means clustering algorithm will be compared.

VI.CONCLUSION

The result analysis will show K-means algorithm performance with different distance measure. Our research will compare different K-means algorithm distance measures. K-Means clustering makes the distances of objects in the same cluster as small as possible, but another objective that the distances of objects from different clusters is not taken into account. This paper will present comparative performance analysis how these two objectives will be satisfied. Different cost function for weighting measure for k-means clustering algorithm will be compared in our research.

REFERENCES

[1] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," *Morgan Kaufmann Publishers*, San Francisco, 2000.

[2] Taoying Li, Yan Chen: "An Improved k-means Algorithm for Clustering Using Entropy Weighting Measures" Dalian Maritime University, Proceedings of the 7th World Congress on Intelligent Control and Automation June 25 - 27, 2008

[3] Taoying Li, Yan Chen: " Weight Entropy k-means Algorithm for Clustering Dataset with Mixed Numeric and Categorical Data", Dalian Maritime University, Dalian 116026, China, Fifth International Conference on Fuzzy Systems and Knowledge Discovery

[4] Pallavi, Sunila Godara:" An Improved Clustering Approach On Time Series Data Set", RTMC, 2011