International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 1, Jan-Feb 2015

RESEARCH ARTICLE OPEN ACCESS

Contrasting Different Distance Functions Using K-Means Algorithm

Kanika[1], Gargi Narula[2]
Department of Computer Science and Engineering,
Swami Vivekanand Institute of Engineering & Technology, Banur
Punjab - India

ABSTRACT

Data mining is the process of semi-automatically analyzing large databases to find useful patterns where the Prediction based on past history. Some of the prediction mechanisms includes Classification, Regression, Clustering, and Association. Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure

Keywords:- Data mining, Cluster, k-means

I. CLUSTER ANALYSIS

A cluster is a collection of data objects which are similar to one another within the same cluster and dissimilar to the objects in the other clusters. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects.[1,2] There is no exact definition regarding the similarity or dissimilarity between data objects in a cluster. It is generally expressed in terms of distance functions. It is the commonest form of unsupervised learning where learning is done from raw data, as opposed to supervised data where the correct classification of examples is given.[1,2] The major aim of clustering is to group a set of data objects into clusters. A good clustering method will produce high quality clusters having high intra-class similarity and low inter-class similarity.[3]

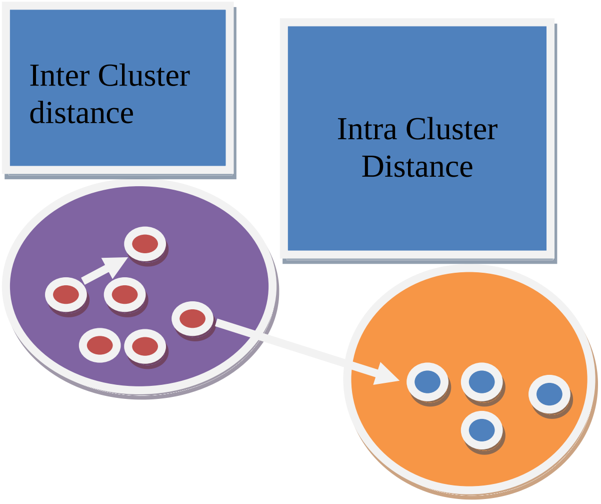

Fig. 1 Inter and Intra Cluster distance

II. SIMILARITY AND DISSIMILARITY BETWEEN OBJECTS

Distances are normally used to measure the similarity or dissimilarity between two data objects.

where i = (xi1, xi2, …, xip) and j = (xj1, xj2, …, xjp) are two p-dimensional data objects, and q is a positive integer

If q = 1, d is Manhattan distance

$$d(i,j) = | x_{i1} - x_{j1} | + | x_{i2} - x_{j2} | + \dots + | x_{in} - x_{jn} |$$

If q = 2, d is Euclidean distance:

d(i,j)=

$$\sqrt{\llbracket(x\rrbracket_{i1} - x_{j1})^2 + \llbracket(x\rrbracket_{i2} - x_{j2})^2 + \cdots + \llbracket(x\rrbracket_{in} - x_{jn})^2}$$ [4]

A. Properties

- d(i,j) ≥ 0
- d(i,i) = 0
- d(i,j) = d(j,i)
- d(i,j) ≤ d(i,k) + d(k,j*)*

III. ALGORITHM USED

K-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group. [5]

The algorithm constitute the following steps:

1. Firstly, the number of clusters must be known, or chosen, to be K say.

2. The initial step is the choose a set of K instances as centres of the clusters. Often chosen such that the points are mutually "farthest apart", in some way.

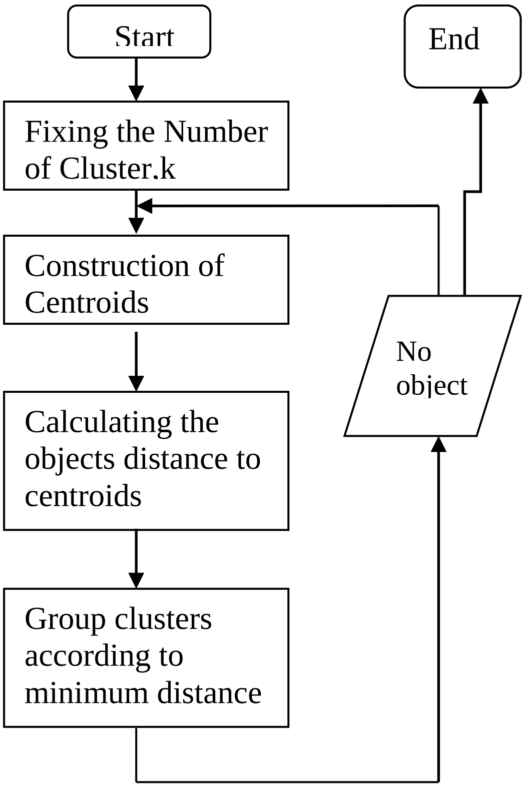
Start
End
Fixing the Number of Cluster.k
Construction of Centroids
Calculating the objects distance to centroids
Group clusters according to minimum distance
No object

Fig. 2 Flow Chart of Algorithm

3. Next, the algorithm considers each instance and assigns it to the cluster which is closest.

4. The cluster centroids are recalculated either after each instance assignment, or after the whole cycle of re-assignments.

5. The steps 3 and 4 are repeated until there are no further centriods to move.

IV. EXPERIMENTAL RESULTS

In this paper, readings are shown which came by applying k-means clustering algorithm on diabetes.arff [6] ,Seed is taken as 10 and maximum iteration is taken as 500 for all above readings. Only distance function is changed with respect to particular value of k for example for k = 2, simple k-means algorithm is run first by taking Euclidean distance function and number of iterations are noted then Manhattan distance

Function taken for k = 2, and number of iterations are again noted down. Similar procedure is repeated for all value of k, number of iterations is noted for both distance functions.

TABLE I
THE EXPERIMENT RESULTS OF NUMBER OF ITERATIONS WITH K-MEANS ALGORITHM USING DIFFERENT DISTANCE

Value of k	NOI in Euclidean distance function	NOI in Manhattan distance function
2	4	6
3	13	21
4	12	16
6	9	11
8	19	18
10	27	11
12	18	20

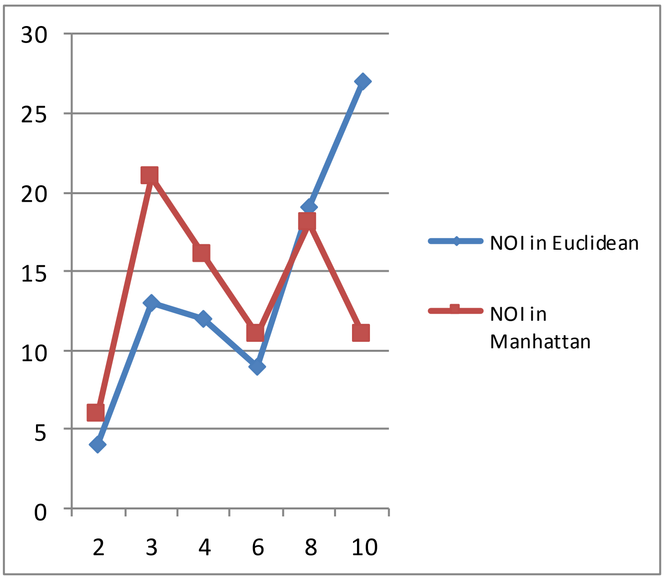

Fig. 3 Plot of number of iterations on X- axis versus Value of K on Y-axis.

The graph shows the result of number of iterations (NOI) comes from using Euclidean distance function and Manhattan distance function with respect to value of k. As seen from the experiment the Manhattan distance function require more number of iterations than Euclidean distance function except when value of k is between 8 to 10. When the value of k=12, the NOI for Euclidean distance function again decreases. As time complexity of k-means algorithm which is O (nkl) where n is the number of patterns, k is the number of clusters and l is the number of iteration describes that time comlexity is directly proportional to the number of iterations. So the time complexity is affected by number of iterations. From the experimnetal results the number of iterations in the Manhattan distance function is genrally more than the Euclidean distance function which shows that Manhattan distance function makes k-means algorithm more computational time complex than Euclidean distance function.

V. CONCLUSION

Clustering is grouping of objects, the objects in a group should be related to one another and unrelated to the objects in other group. k–means clustering algorithm is unsupervised partitioning algorithm which is simple and efficient to implement. This algorithm is classifies the data objects in k different clusters. In k-means clustering algorithm different type of distance functions can be used to measure the distance between two objects. In the experiment Euclidean distance function and Manhattan distance functions are taken to see the effect of these distance function on clustering. As seen from the experiment the Manhattan distance function require more number of iterations than Euclidean distance function except when value of k is between 8 to 10. K-means algorithm is useful for undirected knowledge discovery and is relatively simple. K-means has found wide spread usage in lot of fields, ranging from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others

REFERENCES

[1] Jain A., Murthy M. & Flynn P., "*Data Clustering: A Review*" vol. 31, No. 3, September 1999.

[2] Li C. , Biswas G., "*Unsupervised learning with mixed numeric and nominal data*" vol. 14, No. 4, July/August 2002, pp. 676-69.

[3] Levent Ert¨oz, Michael Steinbach, Vipin Kumar, "*Finding Clusters of Different Sizes, Shapes, and Densities in Noisy,High Dimensional Data*, SIAM International Conference on Data Mining ,February 20, 2003

[4] Suvog Rao, Alfredo Rodriguez, Gary Benson," Evaluating distance functions for clustering tandem repeats, GENOME INFORMATICS SERIES,2005.

[5] K-means, http://en.wikipedia.org/wiki/k-means_clustering

[6] http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff