National Conference on Architecture, Software systems and Green computing-2013(NCASG2013)

A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms

B.S.Charulatha, Paul Rodrigues, T.Chitralekha, Arun Rajaraman Member IEEE

Abstract- Clustering is a process of collecting similar data or objects into groups. The objects are represented in an n-dimensional Euclidean space. Hence the similarity can be brought in by the distance metrics. There are various distance metrics which can be used to find the similarity to form the groups. Fuzzy clustering is a broad classification of clustering methods .They are helpful when there exists a dataset with sub groupings of points having indistinct boundaries and overlap between the clusters. Fuzzy clustering also uses the above said distance metrics for grouping and also uses a membership function which determines by what percentage the object belongs to a group. Although many algorithms exist, the most common is Fuzzy C Means algorithm. By default in FCM the similarity metric used is Euclidean distance. This paper reviews FCM with five distance metrics can be used with fuzzy clustering. They are the Euclid, Manhattan, Canberra, TChebychev and Cosine or angular. Performance of the metrics are presented and compared.

Keywords: Clustering ,Fuzzy Clustering, Fuzzy C Means, Euclidean ,Manhattan, Canberra, TChebychev, Cosine,

I.INTRODUCTION

Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters. Clustering is one of the Data mining techniques. It is an unsupervised learning. Unsupervised learning means there will be no training phase and testing phase instead the learning is by observation. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity.

B.S. Charulatha is with Jawahar Engineering College, India (charu2303@yahoo.co.in)
Paul Rodrigues is with Velammal Engineering College
T.Chitralekha is with Pondicherry University,India
Arun Rajaraman Retired Professor IIT M,India

There are many ways to determine the similarity between two things. In order to represent this similarity in a machine, define a similarity score. If it is possible to quantify different attributes of data objects, different similarity algorithms can be applied across those attributes that will yield similarity scores between the different data objects.

Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed. Some examples of values that can be used as similarity measures include distance, connectivity, and intensity.

The tasks of clustering are

A) Given a set of p variables X_1, X_2,…, X_p, and a set of N objects, the task is to group the objects into classes so that objects within classes are more similar to one another than to members of other classes.

B) Given a set of p variables X_1, X_2,…, X_p, and a set of N objects, the task is to group the variables into classes so that variables within classes are more highly correlated with one another than to members of other classes.

C) Given a set of p variables X_1, X_2,…, X_p, and a set of N objects, the task is to group the objects and variables into classes so that variables and objects within classes are more highly correlated with one another than to members of other classes.

The clustering algorithms can be broadly classified into hard and fuzzy clustering based upon membership. The hard clustering methods restrict that each point of the data set belongs to exactly one cluster. Fuzzy set theory proposed by Zadeh in 1965 gave an idea of uncertainty of belonging which was described by a membership function. The use of fuzzy sets provides imprecise class membership information. Applications of fuzzy set theory in cluster analysis were early proposed in the work of Bellman, Kalaba and Zadeh and Ruspini

ISBN NO: 978-93-80609-14-0

The major differences between hard and fuzzy clustering are as follows:

Hard Clustering

i) Data is divided into crisp clusters, where each data point belongs to exactly one cluster.
ii) Degree of membership is either 0 or 1
iii) Hard clustering method leads to local optimum

Fuzzy Clustering

i) The data points can belong to more than one cluster, based on membership function.
ii) Degree of membership is between 0 and 1.
iii) Fuzzy method leads to global optimum

Clustering adds to the value of existing databases by revealing hidden relationships in the data, which are useful for understanding trends, making predictions of future events from historical data, or synthesizing data records into meaningful clusters.

The application of clustering is spread out in various areas such as medicine, geology, business, engineering systems and image processing, etc., is well documented in the literature.[1,2]

The remainder of the paper is organized as follows. Section II presents an overview of data clustering techniques and the underlying concepts. Section III presents the FCM algorithm in detail along with the underlying mathematical foundations. Section IV introduces the implementation of the techniques and Section V goes over the results of each distance metric followed by a comparison of the results. Section VI concludes the paper.

II CLUSTERING TECHNIQUES

Table I list the five main clustering paradigms. The table describes the main feature of each paradigm. Each of these paradigms is not exclusive and considerable overlap exists between them.[3]

TABLE I

Paradigms	**Description**
Hierarchial	Produces a tree-like description of the clustering structure.The tree is constructed by recursively merging similar objects to form clusters, then merging the clusters to form new super-clusters, this ends when all clusters have merged into one super-cluster. Cutting the tree at any level provides a partition of the objects. [Bajcsy & Ahuja, 1998], [ElSonbaty & Ismail, 1998]
Graph-theoretic	Views the objects as nodes in a weighted network, or graph. This is very helpful for two-dimensional dot patterns. The weight between one node and another is the distance between them using an appropriate metric. The problem, thus, becomes a graph-theoretic one where, for example, a minimal spanning tree is constructed on the dot pattern. This can help illustrate the clustering structure. [Brito *et al.*, 1997], [Pacheco, 1998], [Shapiro, 1995]
Mixture Models	Assumes the objects were generated by a mixture of probability distributions. Determination of the parameters of each distribution, defines the clusters. [McLachlan & Basford, 1988], [Fraley & Raftery, 1998], [Banfield & Raftery, 1993]
Partitional	Clusters are disjoint partition of objects. An object belongs to only one cluster; crisp membership. Usually employs notion of prototypes around which objects cluster, and an objective function to assess a given partition. [Lin & Lin, 1996], [AlSultan & Khan, 1996]
Fuzzy	An object possesses varying degrees of membership with more than one cluster. Extends partitional paradigm, but extensions for all other paradigms are being proposed. [Bezdek, 1981], [Hoppner *et al.*, 1999], this dissertation

Paradigms of Clustering

ISBN NO: 978-93-80609-14-0

Partition Clustering

This paper focuses on Partitioning Clustering paradigms

In partitioned clustering, the object is to partition a set of *N* objects into a number k predetermined clusters by maximizing the distance between cluster centers while minimizing the within-cluster variation.

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

Step 1 Make the first object the centroid for the first cluster.

Step 2 For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

KMeans Algorithm

A method of partitioned clustering whereby a set of k clusters is produced by minimizing the cost based on Euclidean distances. This is very much like a single-classification of groups, except that groups are not known a priori. Because k-means clustering does not search though every possible partitioning, it is always possible that there are other solutions yielding smaller cost

FCM

Fuzzy C-means clustering, which was proposed by Bezdek in 1973 is an improvement over Hard C- means clustering. In this technique each data point belongs to a cluster to a degree specified by a membership grade. As in K-means clustering, Fuzzy C-means clustering relies on minimizing a cost function of dissimilarity measure.

III FUZZY C MEANS

Fuzzy C-means clustering (FCM), relies on the basic idea of Hard C-means clustering (HCM), with the difference that in FCM each data point belongs to a cluster to a degree of membership grade, while in HCM every data point either belongs to a certain cluster or not. So FCM employs fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness specified by membership grades between 0 and 1. However, FCM still uses a cost function that is to be minimized while trying to partition the data set.

The membership matrix **U** is allowed to have elements with values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity:

$\sum_{i=1}^{c} u_{ij} = 1, \forall_j = 1,2,3 \ldots, n$ Eqn (1)

The cost function for FCM is:

$J(U,c_1,c_2,\ldots c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2}$ Eqn2

Where u_{ij} is between 0 and 1; c_i is the cluster center of fuzzy group i; d_{ij} is the Euclidean distance between tne ith cluster centre and jth data point

M ε[1,∞] is a weighting exponent

The necessary condition for Eqn (2) to reach its minimum are Eqn 3 and 4

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \quad \text{Eqn 3}$$

$$u_{ij} = \frac{1}{\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \quad \text{Eqn 4}$$

The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. In a batch mode FCM determines the cluster centers $\mathbf{c}_i$ and the membership matrix **U** using the following steps:

Step 1: Initialize the membership matrix **U** with random values between 0 and 1 such that the constraints in Equation (1) are satisfied.

Step 2: Calculate *c* fuzzy cluster centers c_i , 1, , ,c using Equation (3).

Step 3: Compute the cost function according to Equation (2). Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

The performance of FCM depends on the initial membership matrix values; thereby it is advisable to run the algorithm for several times, each starting with different values of membership grades of data points.

IV METRICS TO FIND THE SIMILARITY

Clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar is required. There are two main type of measures used to estimate this relation: distance measures and similarity measures. Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects.

Distance Metrics

The four distance metrics and one similarity metrics discussed in this paper are [7,8]

1 Euclidean Distance
2 Manhattan
3 Canberra
4 Tchebychev

And Similarity metric is Cosine Metric

Euclidean Distance

The Euclidean distance or Euclidean metric is the ordinary distance between two points that one would measure with a ruler. It is the straight line distance between two points. In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is v((x1 - x2)² + (y1 - y2)²). The distance is calculated using the formula

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{Eqn 5}$$

Manhattan Distance

The distance between two points measured along axes at right angles. In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is |x1 - x2| + |y1 - y2|.The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components. The formula for this distance between a point *X*=(*X1*, *X2*, etc.) and a point *Y*=(*Y1*, *Y2*, etc.) is:

$$d = \sum_{i=1}^{n}|x_i - y_i| \quad \text{Eqn 6}$$

Chebychev Distance

The Chebychev distance between two points is the maximum distance between the points in any single dimension. The distance between points *X*=(*X1*, *X2*, etc.) and *Y*=(*Y1*, *Y2*, etc.) is computed using the formula:

$$max_i = |x_i - y_i| \quad \text{Eqn 7}$$

where *Xi* and *Yi* are the values of the *i*th variable at points *X* and *Y*, respectively. The Chebychev distance may be appropriate if the difference between points is reflected more by differences in individual dimensions rather than all the dimensions considered together. This distance measurement is very sensitive to outlying measurements.

Canberra Distance Metric

The Canberra distance is a metric used for data scattered around the origin.It was introduced in 1966.The formula to calculate the distance is

$$d(i,j = \sum_{k=1}^{n}\frac{|y_{ik}-y_{jk}|}{|y_{ik}|+|y_{jk}|} \quad \text{Eqn 8}$$

Cosine Similarity Metric

Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. The result of the Cosine function is equal to 1 when the angle is 0, and it is less than 1 when the angle is of any other value. As the angle between the vectors shortens, the cosine angle approaches 1, meaning that the two vectors are getting closer, meaning that the similarity of whatever is represented by the vectors increases.

$$s(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad \text{Eqn 9}$$

where s is the similarity metric.

V EVALUATION

The number of clusters into which the data set is to be partitioned is two clusters. The data set is an image with 4096 X 16 [6]. As mentioned earlier, the similarity metric used to calculate the similarity between an input vector and a cluster center is the different distance metrics. Since most similarity metrics are sensitive to the large ranges of elements in the input vectors, each of the input variables must be normalized to within the unit interval [0,1] the data set has to be normalized to be within the unit hypercube.

FCM allows for data points to have different degrees of membership to each of the clusters This approach employs fuzzy measures as the basis for membership matrix calculation and for cluster centers

identification. FCM starts by assigning random values to the membership matrix **U** , thus several runs have to be conducted to have higher probability of getting good performance.

Table 2 lists the results of the tests with the effect of varying the weighting exponent *m* . The experiment was done in MATLAB 7.9.0. The cell value is the iteration number it takes to converge.
The value of 2 seems adequate for this problem since crisp values (either 1 or 0), the evaluation set degrees of membership are defuzzified to be tested against the actual outputs. The graph below is the distance metrics vs number of iterations

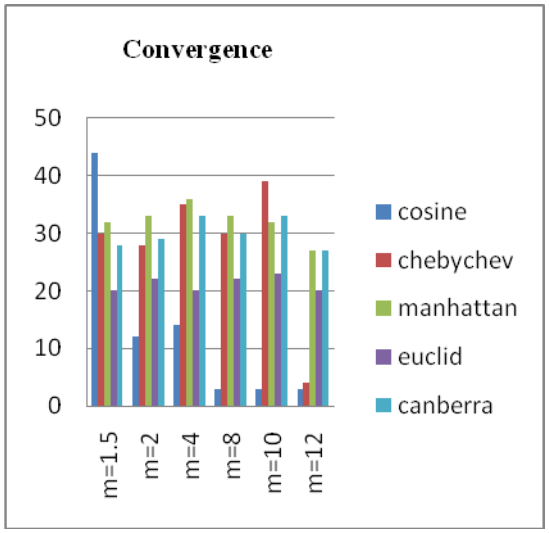

Fig 1 Graph showing the iterations for different values of m

Table 2

	m=1.5	m=2	m=4	m=8	m=10	M=12
cosine	44	12	14	3	3	3
chebychev	30	28	35	30	39	4
manhattan	32	33	36	33	32	27
euclid	20	22	20	22	23	20
canberra	28	29	33	30	33	27

Iterations needed to converge

VI CONCULSION

The paper concludes with the comparison among the above said five metrics. This is only a sample study since the similarity metrics are many. Exhaustive exploration on the distance metrics needs to be done on various data sets on various clustering algorithms.

REFERENCES

[1] B.S.Charulatha,PaulRodrigues,T.Chitralekha Fuzzy clustering algorithms -Different methodologies and parameters- A survey ,International Conference on Advances in Electrical and Electronics, Information Communication and Bio Informatics 2012.
[2] M.-S. Yang-A Survey Of Fuzzy Clustering Mathl. Comput. Modelling Vol. 18, No. 11, Pp. 1-16, 1993
[3] Ahmed Ismail Shihab-Fuzzy Clustering Algorithms And Their Application To Medical Image Analysis,Ph.D Dissertion,University Of London,2000
[4] Khaled Hammouda Prof. Fakhreddine Karray A Comparative Study of Data Clustering Techniques
[5] Lior Rokach, ,Oded Maimon,Clustering Methods,Department of Industrial Engineering Tel-Aviv University
[6] http://cs.joensuu.fi/sipu/datasets/bridge.txt
[7] Shraddha Pandit, Suchita Gupta , A Comparative Study On Distance Measuringapproaches For Clustering, International Journal of Research in Computer Science eISSN 2249-8265 Volume 2 Issue 1 (2011) pp. 29-31 © White Globe Publications
[8] Wikipedia for distance metrics

ISBN NO: 978-93-80609-14-0