ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 72-74 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

Analyzing Formation Of K Mean Clusters Using Similarity And Dissimilarity Measures

Kompal Ahuja[1], Ajmer Saini[2]
[1] Computer Science & Engineering, DCRUST ,Murthal, Haryana, India, kanuahuja@yahoo.co.in
[2] Computer Science & Engineering, DCRUST ,Murthal, Haryana, India, ajmer.saini@gmail.com

Abstract *:* Measurement of similarity and dissimilarity between two data objects is a challenging problem for data mining **.** Mainly the clustering algorithms use distance function between the data points to define dissimilarity measure to form clusters but normalized scaler product can also be used as similarity function to form clusters among data points. This paper presents the results of an experimental study of some common k mean clustering techniques. In particular, we compare the two main approaches to k mean clustering. (we used Euclidean distance as dissimilarity measure and cosine similarity .)

Key words: k mean algorithm, Euclidean distance ,cosine similarity .

INTRODUCTION

Data clustering has been used as the hot topic for research in the field of data mining. The clustering techniques are applied and used to perform search, pattern recognition, trend analysis and so forth on the basis of similarity and dissimilarity measures[2] .Clustering is the technique of grouping a set of physical or abstract objects into different clusters, such that objects with in a cluster are more similar to one another and are dissimilar to the objects in other clusters. A good clustering algorithm generates high quality clusters to yield low inter cluster similarity and high intra cluster similarity.

The knowledge about similarity and dissimilarity is necessary for data mining, pattern recognition, machine intelligent, artificial intelligent and multi-agents system fields[3]. Popularity of this application is not only limited to computer science but the Other fields of natural and social science as well as engineering and statistics are also influenced by it . Data mining tools such as K means clustering rely heavily on the distance to find the similarity between datasets.

Similarity is the quantity that reflects the strength of relationship between two objects. By finding out the similarity and dissimilarity between the objects

(1)we can distinguish one object from another

(2)We can group them (for example using k-means clustering).and can understand the characteristics of each group.

(3)We can explain the behavior of the clusters.

(4)Grouping also may give more efficient organization and retrieval of information.

(5)We can classify a new object into the group

(6)We can predict the behavior of the new object.

(7) We can take action, plan and decision based on the structure and prediction of the data.

Similarity and dissimilarity can be measured for two objects based on several features variables. Depending on the measurement scale of the features variable, similarity and dissimilarity can be determined. After the similarity/dissimilarity of each variable is determined, we can aggregate all features variables together into single Similarity/ dissimilarity index between the two objects[3].

K-MEANS ALGORITHM

K-means clustering is a partitioning method. The function `kmeans` partitions data into *k* mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation. Unlike hierarchical clustering, *k*-means clustering creates a single level of clusters. The distinctions mean that *k*-means clustering is often more suitable than hierarchical clustering for large amounts of data.

`K means` treats each observation in your data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster in the partition is defined by its member objects and by its centroid , or center. The centroid for each cluster is the point to which the sum of dissimilarity from all objects in that cluster is minimized and the similarity measure is maximised .

Suppose a dataset D, contains n objects in Euclidean space . partitioning methods distribute the objects in D into k clusters , $C_1,\ldots..C_k$, that is C_i (D and $C_i \cap D$ and $C_i \cap C_j = \emptyset$ for $(1 \leq i, j \leq k)$)

The algorithm accepts two inputs. The data itself, and "k", the number of clusters. We will talk about the implications of specifying "k" later. The output is k clusters with input data partitioned among them.

The aim of K-means (or clustering) is : to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is ,the objective function aims for high inter

cluster similarity and low inter cluster similarity. We use the distance measures and normalized scaler product to calculate dissimilarity and similarity respectivily .

Algorithm

1.Randomly choose k items and make them as initial centroids.

2. For each point, find the nearest centroid and assign the point to the cluster associated with the nearest centroid.

3. Update the centroid of each cluster based on the items in that cluster. Typically, the new centroid will be the average of

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 72-74 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

all points in the cluster.
4. Repeats steps 2 and 3, till no point switches clusters

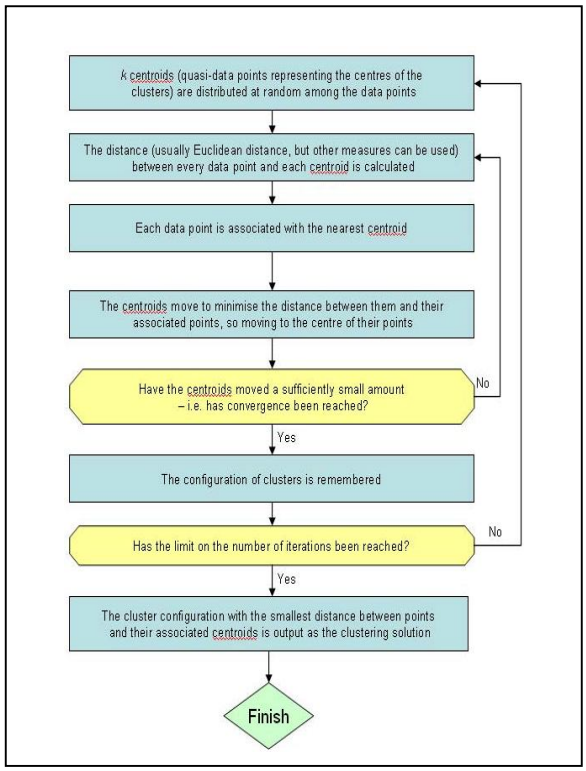

Fig 1: shows flowchart for k-mean clustering using Euclidean distance

Working of k mean

The k mean algorithm defines the centroid of a cluster as the mean value of the points within the cluster .it randomly selects k of the objects in D ,each of which initially represents a cluster mean or center .For each of the remaining objects , an object is assigned to the cluster to which it is most similar ,based on the Euclidean distance / cosine similarity between the object and the cluster mean .the k mean algorithm then iteratively improves the within cluster variation .For each cluster ,it computes the new mean using the objects assigned to the cluster in the previous iteration .All the objects are then reassigned using the updated mean as new cluster centers. The iteration continous until the assignment is stable. [5]

DISSIMILARITY AND SIMILLARITY : DISTANCE MEASURES

An important step in any clustering is to select a distance measure because if we want to compare two items we need a notion of similarity or dissimilarity. One of the common way to find the dissimilarity is to find the distance between the objects. lesser the distance lesser dissimilar are the objects. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.

In 2dimensional space ,the distance between the point (x=1,y=0) and the origin (x=0,y=0) is always one according to the usual norms ,but the distance between the point (x=1,y=1) and the origin can be 2,$\sqrt{2}$,or 1 if you take respectively the 1 norm,2 norm or infinity norm distance

There many distance measures like euclidean distance, Manhattan Distance , Mahalanobis distance..Things are a bit complicated when the items are not points but objects with some attributes (for eg news articles). In this we might need to design our own distance measures. For checking if news articles talk about same topic, a simple distance measure can be to find the entities in an article and compare their frequencies. Two articles with common words like India, Bangladesh, Cricket, Victor, Test "might" talk about the same topic[3]. For a real world application , finding a good distance measure requires lot of work – The distance measure should be efficient but also reasonably accurate.

Euclidean distance

Euclidean distance is the most popular distance function .It is also called as “straight line” or “as crow flies” or 2-norm distance. A review of cluster analysis in health psychology found that the most common distance measure in publish studies in that research area is the Euclidean distance. Let ,

$i= (x_{i1},x_{i2},\ldots..,x_{ip})$ and $j =(x_{j1},x_{j2},\ldots..,x_{jp})$

be two objects described by p numeric attributes .The Euclidean distance between objects i and j is defined as

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})+(x_{i2}-x_{j2})+\cdots+(x_{ip}-x_{jp})}$$

Euclidean distance satisfies following properties:-

1. **Non –negativity** - $d(i,j) \geq 0$: Distance is the non-negative number.
2. **Identity of indiscernibles**: $d(i,i) = 0$.Thedistance of an object to itself is 0.
3.**Symmetry**:$d(i,j)=d(j,i)$:distance is symmetric function .
4.**Triangle inequality** : $d(i,j) \leq d(i,k) + d(k,j)$.

A measure that satisfies that satisfies these conditions is called metric .non negativity property is implied by the other three properties[5].

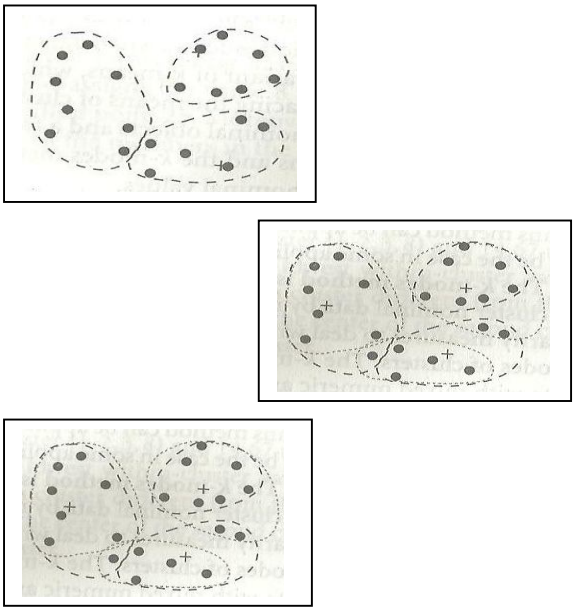

Fig 2: shows formation of clusters

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 72-74 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

Cosine similarity

Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. This rule is particularly useful when clusters develop along certain principal and different axes .

$$\mathbf{Cos}\ \varphi = \frac{x \cdot x_1}{\|x\| \cdot \|x_1\|}$$

For $\cos\varphi_2 < \cos\varphi_1$,pattern x is more similar to $x_2$than x_1.It would be apparent to group it with the second of the two apparent clusters To facilitate this decision ,the threshold angle φ_t can be chosen to define the minimum angular cluster distance. The cosine of 0 is 1, and less than 1 for any other angle; the lowest value of the cosine is -1. The cosine of the angle between two vectors thus determines whether two vectors are pointing in roughly the same direction.

Properties of cosine similarity:-

1.$sim(x_i,x_j)=0$ if xi and xj are not alike at all
2.$sim(x_i,x_j)=1$
3.$sim(x_i,x_j) < sim(x_i,x_k)$ if xi is more like x_k than it is like x_j

PROPOSED WORK

In future we will be formulating an guideline system in limelight of similarity and dissimilarity techniques which are discussed above ,which would be a boom to the all industries.Development of guideline system , will help the customers at the point of care and controlling of different pathways. A person can determine efficient way to organize and manage things. This project easily predicts the condition and different stages of production etc on the basis of the indication provided to system. When we provide temporal data about product or customer so, we can easily detect what will be the condition of product or customer in future and giving appropriate guidelines for it.which will save time and effort (repeatition of same treatment can be avoided easily).This proposed system will
1 Extract the hidden knowledge present in databases.
2Prioritizing these evidences according to quality: Divide the result in two parts Rejected Cases and Accepted Case. For Accepted Cases give guideline for future. For Rejected Cases send for further evaluating. Now Accepted Cases became " Evidences".

CONCLUSION

In this paper, we made a study on similarity and dissimilarity measures namely used in k-means clustering algorithm. We analyzed that both the measures of k mean clustering can be used in guideline systems for taking appropriate decisions in absence as well as presence of supervisor. Data mining supportive systems will provide quality of services whether they are diagnosing correctly or administrating things which are effective .Using data mining techniques in industry one can be sure about saving its time .They are reliable, powerful ,user friendly platform for strategic decision making .Such measures ensures avoidance of duplications of tasks .However supervised learning methods provide desired output sets, but inturn occupies a tremendous database space also it requires trained dataset,which makes its application limited.

REFERENCES

[1]International Journal of Computer Applications (0975 – 8887)Volume 17– No.1, March 2011,25
A Novel Similarity Measure for Clustering CategoricalData Sets Rishi Sayal ,HOD & Professor, Dept. of Computer Science &Engineering, Guru Nanak Engineering College, Ibrahimpatnam,Andhra Pradesh, India.Dr. V. Vijay Kumar,Dean & Professor, Dept. ofComputer Science & Engineering,GIET, Rajahmundry, Andhra Pradesh, India.
[2] Rui Xu, Donald Wunsch II, "*Survey of Clustering Algorithms", IEEE in Neural Networks* 16(3)(2005).
[3]www.karditeknomo.com
[4]www.springerimages.com
[5]*data mining concepts and techniques* by jiawei han |micheline kamber |jian pei
[6] *An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm*
Dost Muhammad Khan, Nawaz Mohamudally
[7] *Data mining introductory and advanced topics* by Marget H. dunhum S.sridhar
[8] *Introduction to Artifiial Neural Systems*

.