Effect of Distance Functions on K-Means Clustering Algorithm

Richa Loohach
Research Scholar, Dept. Of Computer Science and Applications, Kurukshetra University, Kurukshetra

Kanwal Garg
Phd,Assistant Professor, Dept. Of Computer Science and Applications, Kurukshetra University, Kurukshetra

ABSTRACT
Clustering analysis is the most significant step in data mining. This paper discusses the k-means clustering algorithm and various distance functions used in k-means clustering algorithm such as Euclidean distance function and Manhattan distance function. Experimental results are shown to observe the effect of Manhattan distance function and Euclidean distance function on k-means clustering algorithm. These results also show that distance functions furthermore affect the size of clusters formed by the k-means clustering algorithm.

Keywords
K-means clustering, distance functions, clustering, Euclidean distance function, Manhattan distance function.

1. INTRODUCTION
Clustering is a technique that classifies the raw data reasonably and searches the hidden patterns that may be present in datasets [11]. It is a process of grouping data objects into disjointed clusters; these obtained clusters should reflect some mechanism at work in the domain from which instances or data points are drawn [10], a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances [6]. The Greater the similarity (or homogeneity) within a group and greater the difference between groups, the better is the clustering [7]. There are various techniques available for clustering like k-means clustering technique, hierarchical clustering technique, density based clustering techniques etc. but k-means clustering algorithm is most widely used algorithm because it is simple, efficient and easy to implement [6]. So the author studies the simple k-means algorithm and effect of various distance functions on it because distance function plays major role in finding relationship between various objects in a dataset. K-means is a numerical, non-deterministic, iterative method [2]. So for many practical applications, this method is proved to be very effective to obtain good clustering results. In this paper author first gives the introduction then discussion about k-means clustering algorithm and various distance functions is done in second and third section. In fourth section experimental results are shown and in last section conclusion is given.

2. THE K-MEANS CLUSTERING ALGORITHM
K-means clustering algorithm is first proposed by Macqueen in 1967 which was uncomplicated, non-supervised learning clustering algorithm [9]. K-means is a partitioning clustering algorithm, this technique is used to classify the given data objects into k different clusters through the iterative method, which tends to converge to a local minimum. So the outcomes of generated clusters are dense and independent of each other [12]. The algorithm consists of two separate phases.

(i). In the first phase user selects k centres randomly, where the value k is fixed in advance. To take each data object to the nearest centre. Several distance functions are considered to determine the distance between each data object and the cluster centres. When all the data objects are included in some clusters, the first step is completed and an early grouping is done.

(ii). Then the second phase is to recalculate the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum. The process of k-means algorithm as follow:

Input:
Number of desired clusters value of k (e.g. 2, 3, 4….etc.), and a database D = $\{d_1, d_2 \ldots d_n\}$ containing n data objects.

Output:
A set of k clusters

Steps:

1. Randomly select k data objects from dataset D as initial cluster centres.

2. Repeat step 3 and 4 until there is no change in the centre of clusters

3. Calculate the distance between each data object d_i ($1 <= i <= n$) and all k cluster centres c_j ($1<=j<=k$) and assign data object d_i to the nearest cluster.

4. For each cluster j ($1<=j<=k$), recalculate the cluster centre.

Before the k-means algorithm converges, calculations of distance and cluster centres are done while loops are executed a number of times, where the positive integer t is known as the number of k-means iterations. The accurate value of t varies which mainly depends on the initial cluster centres. The allocation of data points is related to the new clustering centre, so the computational time complexity of the k-means algorithm is *O(nkt)*.where n is the number of all data objects, k is the number of clusters, t is the iterations of algorithm. Usually requiring $k << n$ and $t << n$. Computational time complexity also depends on the complexity of distance function used to measure the distance between the objects of databases[7], the distance is calculated from data object x to each cluster centre and then find that the distance to the cluster C is the smallest. So in the course of several iterations, k-means algorithm is to calculate the distance between data object x to the other cluster centre, which takes up a long execution time thus affecting the efficiency of clustering [1]. Major drawbacks of k-means clustering algorithm are first, it can be actually slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this process is really sensitive to the provided initial clusters, another drawback is that it only covers numerical attributes; also this algorithm lacks scalability and is insensitive with

respect to the outliers [8]. In next section the author discuss about various distance functions used in k means clustering algorithm.

3. VARIOUS DISTANCE FUNCTIONS

Distance functions in k-means clustering technique plays an important role. Different distance functions are provided to measure the distance between data objects. These two distance functions are discussed as follows

3.1. Euclidean Distance Function

Euclidean distance is ordinary distance between two points that one would measure with a ruler. It is the most commonly used distance function [5]. This distance is given by Pythagorean formula. The Euclidean distance between the points a and b is the length of the line segment connecting them (a, b) [4]. In the Euclidean plane, if a = (a_1, a_2) and b = (b_1, b_2) then the distance is given by:

D (a, b) = $\sqrt{(a_1-b_1)^2 + (a_2-b_2)^2}$. This is equivalent to Pythagorean formula. Weakness of the basic Euclidean distance function is that if one of the input attributes has a relatively large range, then it can overpower the other attributes [4]

3.2. Manhattan distance function

In Manhattan distance function the distance between two points is the sum of the absolute differences of their coordinates. The Manhattan distance, D_1 between two vectors a ,b in an *n*-dimensional real vector space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axis [5]. More formally,

D_1 (a, b) = $\|a-b\|_1 = \sum_{i=1}^{n} |a_i-b_i|$, where a = ($a_1$, a_2… a_n) and b = (b_1, b_2… b_n) are vectors

4. EXPERIMENTAL RESULTS

In this paper WEKA 3.6.5 version software for data mining is used and cpu.arff dataset is used for experimentation which can be obtained from UCI machine learning repository [3]. Author run the software for clustering experiment using k means algorithm and see the different results comes by using different distance functions. Data set cpu.arff contains 209 instances and 7 attributes. NOI means number of iterations.

Table 1. The Experimental Results with K-Means Algorithm (NOI - no. of iterations)

Value of K	NOI in Euclidean distance function	NOI in Manhattan distance function
2	12	7
3	10	8
4	10	13
5	8	10
6	8	7
7	12	14
8	11	11
9	11	8
10	12	10
11	13	6
12	12	6
13	12	8
14	10	9
15	10	7
16	10	7
17	10	6
18	8	6
19	14	13
20	7	7
21	8	7
22	11	7
23	11	7
24	9	7
25	9	7

In above table, readings are shown which came by applying k-means clustering algorithm on cpu.arff dataset. 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 is taken for k = 2, and number of iterations are again noted down. Similar procedure is repeated for all values of k from 2 to 25, number of iterations is noted for both distance functions.

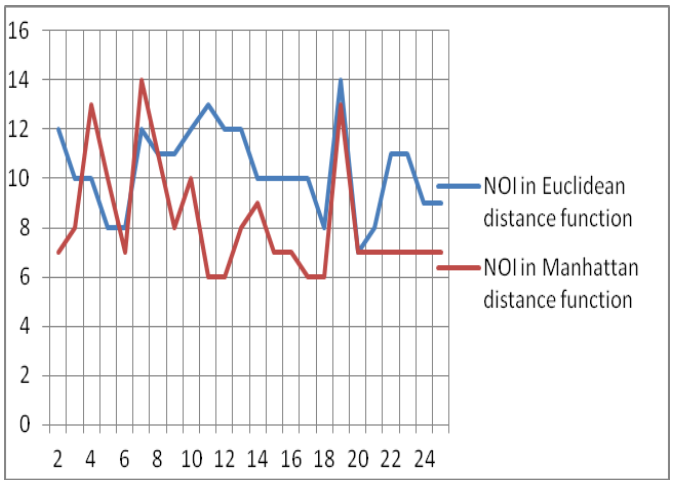

Figure 1: Effect of distance functions on k means clustering, X axis shows the value of k and Y axis represents the number of iterations

The figure 1 shows the result of number of iterations (NOI) comes from using Euclidean distance function and Manhattan distance function with respect to the value of k. As seen from the experiment that Euclidean distance function require more number of iterations than Manhattan distance function except when value of k= 4, 5, 7. Manhattan distance and Euclidean function has same number of iteration at k = 8 and 20. As according to the computational time complexity of k-means algorithm which is O (nkt) suggest that time complexity is directly proportional to the number of iterations [7]. So the computational time complexity is affected by number of iterations. From the experimental results the number of iterations in the Euclidean distance function is generally more than the Manhattan distance function which shows that Manhattan distance function makes k-means algorithm less computational time complex than Euclidean distance function. Figure 1 shows that the efficiency of k-means is more when using Manhattan distance function up to k=4 then it decreases and at k=6 it again increases but it decreases for k=7 after that its efficiency is more than Euclidean distance function continuously up to k =25 There is also difference in the

clusters like with k=2 using Euclidean distance function number of objects in cluster 0 is 171 and in cluster 1 is 38 whereas for the same case when Manhattan distance function is used instead of Euclidean distance function then number of objects in cluster 0 is 146 and in cluster 1 is 63, means members of clusters are different with different distance function which specifies that different distance functions also influence the members of a cluster.

5. CONCLUSION

Clustering is an NP hard problem of grouping of objects, the objects in a group should be related to one another and unrelated to the objects in other groups [6]. K means clustering algorithm is unsupervised partitioning algorithm which is simple and efficient to implement [6]. This algorithm classifies the data objects in k different clusters. In k-means clustering algorithm different types 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 Manhattan distance function outperform the Euclidean distance function from k=2 to 25 except on value k=4, 5, 7 where Euclidean distance function perform better because the number of iteration in Manhattan distance function is less than the Euclidean distance function. This is because computational time complexity is directly proportional to the number of iterations. Also the efficiency of k-means clustering is increases when Manhattan function is used as seen from figure 1 except with k= 4, 5, 7 where Euclidean distance function have better efficiency. Whereas both distance function have same efficiency for k= 8 and 20. Furthermore Manhattan distance function requires less computation than Euclidean distance function which in turn improves the computational time complexity of k-means [4]. Additionally Distance functions also affect the size and members of a cluster as different distance functions uses different approach to find the distance between the data objects which is the most important step of creation of clusters. So distance functions should be chosen wisely and according to the dataset.

6. REFERENCES

[1] Shi Na , Liu Xumin and Guan yong 2010 "Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics 978-0-7695-4020-7/10 $26.00 © 2010 IEEE

[2] A.K. Jain, M.N. Murty and P.J. Flynn 1999, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[3] Source: collection of regression datasets by Luis Torgo (ltorgo@ncc.up.pt) at http://www.ncc.up.pt/~ltorgo/Regression/DataSets.html

[4] D. Randall Wilson and Tony R. Martinez 1997 "Improved Heterogeneous Distance Functions" Journal of Artificial Intelligence Research 6 (1997) 1-34 Submitted 5/96; published 1/97 © 1997 AI Access Foundation and Morgan Kaufmann Publishers. All rights reserved

[5] Antoni Moore 2002 "The case for approximate Distance Transforms" Presented at SIRC 2002 – The 14th Annual Colloquium of the Spatial Information Research Centre University of Otago, Dunedin, New Zealand December 3-5th 2002

[6] Glenn Fung 2001, "A Comprehensive Overview of Basic Clustering Algorithms" June 22, 2001

[7] Michael Steinbach , Levent Ertöz and Vipin Kumar, "The Challenges of Clustering High Dimensional Data", Access to computing facilities was provided by AHPCRC and the Minnesota Supercomputing Institute.

[8] Pavel Berkhin, " Survey of Clustering Data Mining Techniques", Accrue Software, Inc. Author's Address: Pavel Berkhin, Accrue Software, 1045 Forest Knoll Dr., San Josh, CA, 95129; e-mail: pavelb@accrue.com

[9] Juanying Xie, Shuai Jiang 2010, "A simple and fast algorithm for global K-means clustering", 2010 Second International Workshop on Education Technology and Computer Science, 978-0-7695-3987-4/10 $26.00 © 2010 IEEE DOI 10.1109/ETCS.2010.347

[10] Ren Jingbiao,Yin Shaohong 2010 "Research and Improvement of Clustering Algorithm in Data Mining", 2010 2nd International Conference on Signal Processing Systems (ICSPS) 978-1-4244-6893-5/$26.00 C 2010 IEEE

[11] H. G. Wilson, B. Boots, and A. A. Millward 2002, "A Comparison of Hierarchical and Partitional Clustering Techniques for Multispectral Image Classification", 0-7803-7536-X/$17.00 (C) 2002 IEEE

[12] Tung-Shou Chen, Tzu-Hsin Tsai, Yi-Tzu Chen, Chin-Chiang Lin, Rong-Chang Chen, Shuan-Yow Li and Hsin-Yi Chen 2005, "A Combined K-Means And Hierarchical Clustering Method For Improving The Clustering Efficiency Of Microarray", Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems December 13-16, 2005 Hong Kong, 0-7803-9266-3/05/$20.00 ©2005 IEEE