IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

A Study of Different Similarity Measures on the Performance of Fuzzy Clustering

O.A. Mohamed Jafar

Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Tamil Nadu, India

e-mail: mdjafar2021@gmail.com

Available online at: www.ijcseonline.org

Received: 28/Mar/2018, Revised: 06/Apr/2018, Accepted: 21/Apr/2018, Published: 30/Apr/2018

Abstract— Data mining is a collection of exploration methods based on advanced analytical tools and techniques for handling huge amount of information. Clustering is a useful technique for discovery of knowledge from a dataset. Distance measure plays an important role in clustering. It is used to measure the similarity or dissimilarity between two data points. Euclidean distance measure is normally used in most clustering methods. Some of the limitations of this measure are inability to handle noise and outlier data points, not suitable for sparse data and clusters with only elliptical shapes. In this paper, fuzzy clustering is proposed using different similarity measures such as non-negative vector similarity coefficient (NVSC), Correlation and Cosine. The performance of the algorithm is compared with various similarity measures using five real life benchmark sets including Wine, Liver Disorders, Pima Indian Diabetes, Haberman's Survival and Statlog (Heart). Experimental results show that fuzzy clustering based on Cosine similarity measure achieves minimum fitness value, minimum intra-cluster distance and maximum inter-cluster distance on various data sets than other similarity measures.

***Keywords*—**Fuzzy Clustering, Similarity Measures, Cluster Validity

I. INTRODUCTION

Data mining is the process of extracting or mining knowledge from large databases. It involves the use of data analysis methods to find previously unknown, meaningful patterns and relationships in huge amount of data. There are several data mining tasks or functionalities including classification, prediction, clustering, regression, time series analysis, summarization, association rules and sequence discovery. The knowledge discovery in databases is iterative and interactive steps including understanding the application domain, extracting the dataset, data pre-processing, data mining, interpretation and discovering knowledge. Clustering techniques partition the objects into groups or clusters based on similarity metrics.

Data clustering is a popular unsupervised classification technique which partitions an unlabelled data set into groups of similar objects. The purpose of cluster analysis is to group sets of data points into classes such that same points are placed in the one cluster while dissimilar points are placed in other clusters. Data clustering is a challenging problem because many factors such as distance measures, criterion functions and initial conditions have come into play in devising a good technique. Clustering of data can be classified into model-based methods, density-based methods, grid-based methods, probabilistic methods, partitioning methods and hierarchical methods.

Model-based methods assume a model for each of the clusters and find the best fit of the data to the given model. They can be either partitioning or hierarchical depending on the structure or model. ***Density-based methods*** cluster objects based on the distance between objects. The data sets can be divided into several subsets according to the density of the data set points. The density is defined as the number of objects in a particular neighbourhood of the data objects. Two points are said to be density-connected if one can go from one point to another through a list of points which have their local density superior to a certain threshold. ***Grid-based methods*** use uniform grid mesh to partition the problem domain into cells. The clustering operations are done on the grid structure. ***Probabilistic methods*** are an attempt to optimize the fit between the data and the model using probabilistic approach. Each cluster can be represented by Poisson, Gaussian or a mixture of these distributions. **Partitioning methods** partition the database into predefined number of clusters. Given a database of 'n' objects, they attempt to determine 'k' groups, which satisfy the following requirements: (1) each object must belong to exactly one group and (2) each group must contain at least one object. ***Hierarchical methods*** create a hierarchical decomposition of the objects. They can be either agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms start with each object forming a separate group. They consecutively merge the data points that are close to one another, until all the groups are merged into one or until the ending condition satisfies. Divisive algorithms begin with the whole set of objects and proceed it into successively smaller groups until each object is in one cluster or a termination condition holds [1] [2].

K-means [3] is an efficient and widely used hard clustering algorithm. It requires the previous knowledge about the number of clusters. Each data object is from only one cluster. K-means model is not suitable for real world data sets in which there are no definite boundaries between the clusters. Fuzzy c-means (FCM) is one of the most important fuzzy clustering methods, initially proposed by Dunn [4] and then generalized by Bezdek [5]. This technique allows one piece of data object to belong to two or more clusters based on degree of membership. The objective of this technique is the assignment of data objects into clusters with varying degrees of membership values. The membership values lie between 0 and 1. The membership value reflects the degree to which the point is more representative of one cluster than the other. Most of the traditional clustering algorithms use Euclidean distance measure. In this distance measure, the points are distributed around the sample average in a spherical manner. It is not good for sparse data. In this paper, a new fuzzy clustering algorithm based on different similarity measures such as NVSC, Correlation and Cosine is proposed. The fuzzy clustering based on Cosine similarity gives better clustering result than other similarity measures in terms of fitness value, intra-cluster distance and inter-cluster distance for five real world data sets from UCI Machine Learning Repository.

This paper is organized as follows. In section II, the clustering problem is described. The related work is given in section III. The methodology is described in section IV. In section V, the experimental results are presented. Finally, section VI concludes the paper.

II. Clustering Problem

Given ‘n’ data objects, allocate each data object to one of ‘k’ clusters such that the sum of squared Euclidean distances between each data object and the centre of its belonging cluster for every such allocated data object is minimized. The mathematical model of data clustering problem is described as follows [6]:

$$Minimize\ \ J(w,c) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \| x_i - c_j \|^2 \qquad (1)$$

$$subject\ \ to\ \ \sum_{j=1}^{k} w_{ij} = 1,\ \ i = 1,2,...,n \qquad (2)$$

$$c_j = \frac{1}{N_j}\sum_{x_i \in C_j} x_i\ \ ,\ \ j = 1,2,...,k \qquad (3)$$

$where\ \ w = w_{ij}$ – association weight of data object x_i with cluster j

$$w_{ij} = 0\ or\ 1 \qquad (4)$$

c – set of k clusters $\{c_1, c_2, ..., c_k\}$
n – number of data objects
k – number of clusters
x_i – location of the i-th data object
c_j – centre of the j-th cluster
N_j – number of data objects belonging to the cluster c_j

III. Related Works

Soumi Ghose and Sanjay Kumar Dubey [7] compared the K-means algorithm with Fuzzy C-Means algorithm. FCM algorithm produced close results to K-means clustering algorithm. Archana Singh et al. [8] implemented the K-means algorithm using three different distance metrics Euclidean, Manhattan and Minkowski. K-means clustering algorithm based on Euclidean distance metric produced better result than other distance metrics. Manhattan distance metrics gave the worst result. Hadi Nasooti et al. [9] proposed K-means Clustering algorithm to evaluate the impact of Euclidean and Manhattan distance metrics using network intrusion detection data. Jasmine Irani et al. [10] prepared a survey on various clustering techniques and similarity measures. They have explained the advantages and limitations of the existing methods. Ms. Kothariya Arzoo and Kirit Rathod [11] implemented K-means Clustering with different distance metrics using geographic data set. V.P. Mahatme et al. [12] presented three different distance metrics Euclidean, Manhattan, Pearson Correlation Coefficient on the performance of K-means and Fuzzy C-means clustering.

IV. Methodology

Many clustering algorithms use distance metric to find the similarity or dissimilarity between pair of objects. There is no distance metric that one can be best applied in all applications. The distance measure has the following properties:

i) Distance is always positive.
ii) Distance from point u to itself is always zero.
iii) Distance from u to v is always the same as v to u.
iv) Distance from point u to point v cannot be greater than the sum of the distance from u to some other point w and distance from w to v.

The Euclidean distance measure is generally applied in most clustering algorithms. Some of its drawbacks are as follows.

i) sensitive to scales of variables involved
ii) difficult to handle noise and outlier data points

iii) distributed around the sample mean in a spherical manner
iv) not suitable for sparse data

In the proposed method, a new fuzzy clustering method using different similarity measures such as Cosine, NVSC and Correlation is implemented. Fuzzy clustering [4][5] permits one piece of data to belong to two or more clusters. Given a data set $X = \{x_1, x_2, ..., x_n\}$, the Fuzzy algorithm partitions a data set into c fuzzy clusters $(2 \le c \le n)$ with $z = \{z_1, z_2, ..., z_c\}$ cluster centroids by minimizing the fitness value. Fuzzy clustering algorithm is widely used in many real world applications.

The cosine distance is generally used as a metric for measuring distance when the magnitude of the vectors does not matter. It is also used to compare the documents in text mining. The similarity between two objects is determined by finding the cosine of angle between the selected any two objects.

Cosine Distance Based Fuzzy Clustering

Step 1: Select the number of clusters c ($2 \le c \le N$); choose fuzziness index m (m>1); initialize the fuzzy partition membership values $U^{(0)}$; iteration error ε =0.00001; Fix the maximum number of iterations *max_it*

Step 2: Set the iteration counter t = 0

Step 3: Calculate the cluster centers z_j , $j = 1, 2, ...c$, using

$$z_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{5}$$

Step 4: Calculate the distance

$$d^2(x, y) = \left(1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \right)^2 \tag{6}$$

Step 5: Calculate the value of the objective function J_m using

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}^m d_{ij}^2 \tag{7}$$

Step 6: Update the fuzzy partition membership values $U^{(t+1)}$ using

$$u_{ij}^{(t+1)} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}^2}{d_{ik}^2} \right)^{\frac{1}{m-1}}} \; ; \; 1 \le i \le n; 1 \le j \le c \tag{8}$$

Step 7: If $\| U^{(t+1)} - U^{(t)} \| < \varepsilon$ or t = *max_it* then stop; otherwise set t = t+1 and go to step 3

NVSC Distance Measure

The NVSC distance measure is calculated by the equation (9)

$$d^2(x, y) = \left(1 - \gamma^2(x, y) \right)^2 \tag{9}$$

where

$$\gamma(x, y) = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)} \tag{10}$$

Correlation Distance Measure

The correlation distance measure is calculated by the equation (10)

$$d^2(x, y) = \left(1 - \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \right)^2 \tag{11}$$

This similarity measure is very useful in some applications include gene expression patterns, proteomics and phenotype analysis. It is a good measure for capturing the similarity of patterns of feature changes. It is bound on the interval [-1,1]. The cosine based fuzzy clustering improves the result than the existing method with Euclidean distance.

V. Experimental Results

The fuzzy clustering is implemented in Java and executed in a Windows 7 Professional OS environment using Core 2 Duo CPU, 2.27 GHs and 4 GB RAM. The fuzziness index, iteration error and maximum number of iteration of fuzzy clustering are 2.0, 0.00001 and 100 respectively.

Data Sets

The performance of fuzzy clustering algorithm is evaluated through five real life benchmark data sets including Wine, Liver Disorders, Pima Indian Diabetes, Haberman's Survival and Statlog (Heart).

Wine Data Set is the result of a chemical analysis of wines grown in the same region in Italy but derived from 3 different cultivars. The analysis concluded the quantities of 13 constituents found in each of the 3 types of wines (class 1: 59 records; class 2: 71 records and class3: 48 records). There are 178 records with 13 numeric attributes. All the attributes are continuous.

Liver Disorders Data Set is created by BUPA Medical Research Ltd. The first five attributes in the data set are all blood tests which are considered to be sensitive to liver disorders that have risen from excessive alcohol consumption. Each line the data file constitutes the record of a single male individual. The data set consists of 345 instances or objects and 2 different types characterized by 6 attributes or features.

Pima Indian Diabetes Data set is allocated to recognize diabetic patients. All patients are females at least 21 years old of Pima Indian heritage. The data set has 768 instances, which are classified into 2 classes (class1: 500 instances, class2: 268 instances). Each instance in this data set has 8 attributes, all numeric-valued. This data set does not have any missing value.

Haberman's Survival Data Set consists of 306 instances with 2 different types characterized by 3 attributes, all numeric-valued. The three attributes are age of patient at the time of operation patient's year of operation and number of positive axillary nodes detected. This data set does not have any missing value. There are 225 instances in the category of patients who survived 5 years or longer (class 1) and 81 instances in the category of patients who died within 5 years (class 2).

Statlog (Heart) Data Set is a heart disease database which consists of 270 instances with 2 different types characterized by 13 attributes. There are two classes in the data set: class 1 (150 instances); class 2: (120 instances). This data set does not have any missing value.

The data sets are normalized and the attribute data are scaled so as to fall within a small specified range, such as 0.0 to 1.0.

Performance Measures

The quality of clustering results is measured using the fitness value and cluster validity measure such as intra-cluster distance and inter-cluster distance [13]. Intra-cluster distance is the average of the sum of all the distances between the objects within a cluster and the centroid of the cluster. The smaller intra-cluster value has the higher quality of clustering. Inter-cluster distance is the sum the distance between all pairs of clusters. The distance between two clusters is defined as the distance between their centroids. The higher inter-cluster value has the good quality of clustering.

Table 1 describes some important performance measures to evaluate the algorithm. The optimum result is also shown the table.

The comparison of fitness value of fuzzy clustering algorithm using different similarity measures is shown in Table 2. The proposed algorithm with cosine similarity measure has the minimum fitness value of 0.361, 0.569, 25.181, 1.699 and 5.982 for Wine, Liver Disorders, Pima Indian Diabetes, Haberman's Survival and Statlog (Heart) respectively. Figures 1 to 5 give the fitness value of fuzzy clustering of different similarity measures for the data sets Wine, Liver Disorders, Pima Indian Diabetes, Haberman's Survival and Statlog (Heart). The intra-cluster distance and inter-cluster distance of different similarity measures are shown in table 3 and table 4 respectively. The proposed algorithm with cosine similarity measure has the minimum intra-cluster distance of 0.0247, 0.0579, 0.0192, 0.0480 and 0.152 for Wine, Liver Disorders, Pima Indian Diabetes, Haberman's Survival and Statlog (Heart) respectively. The algorithm has the maximum inter-cluster distance of 3.0685, 0.3311, 0.4465, 0.4839 and 1.0277 for Wine, Liver Disorders, Pima Indian Diabetes, Haberman's Survival and Statlog (Heart) respectively. The fuzzy clustering based on Cosine similarity measure has the best optimal results in terms of fitness value, intra-cluster distance and inter-cluster distance.

Table 1. Performance Measure with Optimal Result

Performance Measure	**Formula**	**Optimal Result**
Fitness Value	$\sum_{j=1}^{c}\sum_{i=1}^{n} u_{ij}^{m} d_{ij}^{2}$	Minimum
Intra-cluster Distance	$\sum_{i=1}^{k}\sum_{x_i \in C_j} \| x_i - z_j \|$	Minimum
Inter-cluster Distance	$\sum_{i=1}^{k}\sum_{j=i+1}^{k} \| z_i - z_j \|$	Maximum

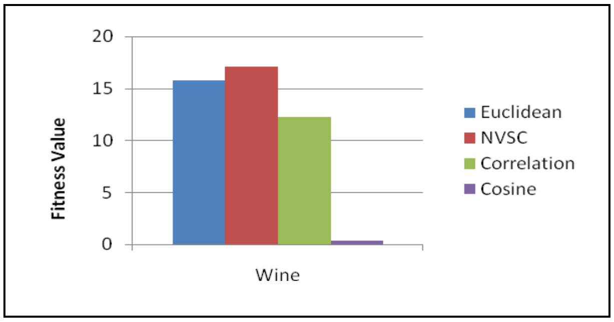

Figure 1. Fitness Value of Wine Data Set

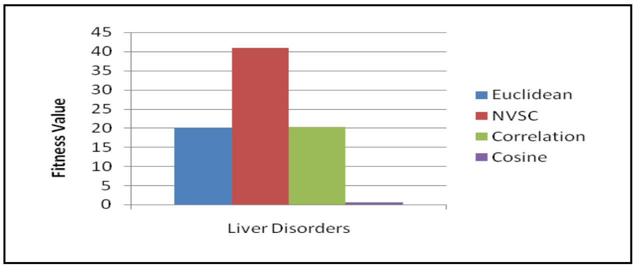

Figure 2. Fitness Value of Liver Disorders Data Set

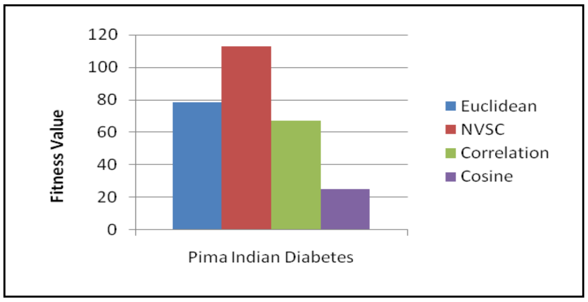

Figure 3. Fitness Value of Pima Indian Diabetes Data Set

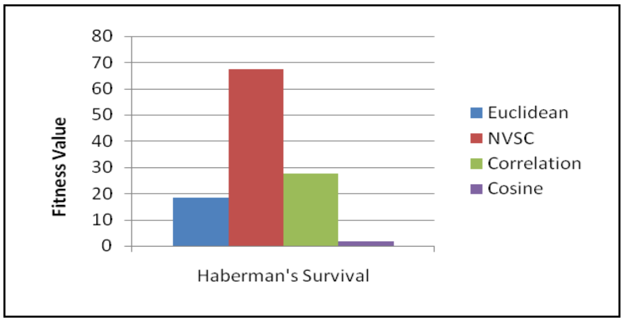

Figure 4. Fitness Value of Haberman's Survival Diabetes Data Set

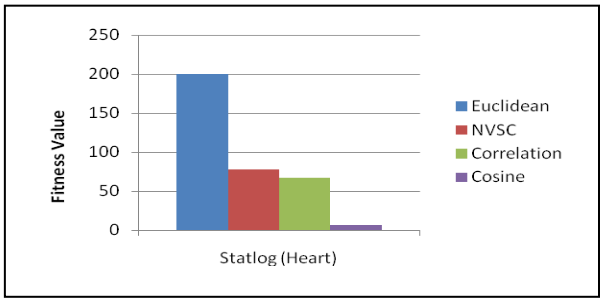

Figure 5. Fitness Value of Statlog (Heart) Data Set

Table 2. Fitness Value Using Different Similarity Measures

Data Set	**Fitness Value**			
	Euclidean	***NVSC***	***Correlation***	***Cosine***
Wine	15.754	17.120	12.216	**0.361**
Liver Disorders	20.050	40.992	20.239	**0.569**
Pima Indian Diabetes	78.300	112.621	67.220	**25.181**
Haberman's Survival	18.481	67.286	27.568	**1.699**
Statlog (Heart)	199.673	77.905	66.708	**5.982**

Table 3. Intra-cluster Distance Using Different Similarity Measures

Data Set	**Intra-cluster Distance**			
	Euclidean	***NVSC***	***Correlation***	***Cosine***
Wine	0.2623	0.6228	0.0622	**0.0247**
Liver Disorders	0.8100	0.7539	0.3989	**0.0579**
Pima Indian Diabetes	0.2159	0.4983	0.0349	**0.0192**
Haberman's Survival	0.6336	0.8997	0.0567	**0.0480**
Statlog (Heart)	1.3034	0.7468	0.5753	**0.152**

Table 4. Inter-cluster Distance Using Different Similarity Measures

Data Set	**Inter-cluster Distance**			
	Euclidean	***NVSC***	***Correlation***	***Cosine***
Wine	0.8066	0.0508	2.9811	**3.0685**
Liver Disorders	0.2499	0.0011	0.2525	**0.3311**
Pima Indian Diabetes	0.2953	0.0001	0.4448	**0.4465**
Haberman's Survival	0.4746	0.1292	0.4804	**0.4839**
Statlog (Heart)	0.5765	0.0001	1.0186	**1.0277**

VI. CONCLUSION

Euclidean distance measure is commonly applied in most clustering algorithms. Some limitations of this distance measure include distributed around the average and not a good measure for sparse data. In this study, Fuzzy clustering is proposed using other similarity measures such as NVSC, Correlation and Cosine. The performance of the algorithm is evaluated through fitness value, intra-cluster and inter-cluster distance using five benchmark real life data sets – Wine, Liver Disorders, Pima Indian Diabetes, Haberman's Survival, Statlog (Heart) from UCI Machine Learning Repository. The proposed algorithm is compared with various similarity measures. The experimental results show that fuzzy clustering based on Cosine similarity measure produces minimum fitness value, minimum intra-cluster distance and maximum inter-cluster distance than other similarity measures.

REFERENCES

[1] J. Han and M. Kamber, *"Data mining: Concepts and Techniques"*, Morgan Kaufmann, San Francisco, 2001.

[2] P. Berkhin, *"Survey clustering Data Mining Techniques"*, Technical Report, Accrue Software, San Jose, California, 2002.

[3] J. MacQueen, *"Some Methods for Classification and Analysis of Multivariate Observations"*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297, 1967.

[4] J.C. Dunn, *"A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters"*, Journal of Cybernetics, Vol. 3, pp. 32-57, 1973.

[5] J.C. Bezdek, *"Pattern Recognition with Fuzzy Objective Function Algorithms"*, Plenum Press, New York, 1981.

[6] J. Dong and M. Qi, *"A new clustering algorithm based on PSO with the jumping mechanism"*, Proceedings of the IEEE third

international symposium on intelligent information technology applications, 2009.

[7] Soumi Ghose and Sanjay Kumar Dubey, *"Comparative Analysis of K-Means and Fuzzy C-Means Algorithms"*, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 4, No. 4, pp. 35-39, 2013.

[8] Archana Singh, Avantika Yadav and Ajay Rana, *"K-means with Three Different Distance Metrics"*, International Journal of Computer Applications, Vol. 67, No. 10, pp. 13-17, 2013.

[9] Hadi Nasooti, Marzieh Ahmadzadeh, Manjeh Kesht Gary and S. Vahid Farrahi, *"The impact of Distance Metrics on K-means Clustering Algorithm Using in Network Intrusion Detection Data"*, International Journal of Computer Networks and Communications Security, Vol. 3, No. 5, pp. 225-228, 2015.

[10] Jasmine Irani, Nitin Pise and Madhura Phatak, *"Clustering Techniques and the Similarity Measures used in Clustering: A Survey"*, International Journal of Computer Applications, Vol. 134, No. 7, pp. 9-14, 2016.

[11] Ms. Kothariya Arzoo and Kirit Rathod, *"K-Means Algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8.2"*, International Research Journal of Engineering and Technology (IRJET), Vol. 4, Issue 4, pp. 2363-2368, 2017.

[12] V.P. Mahatme and Dr. K.K. Bhoyar, *"Impact of Distance Metrics on the Performance of K-Means and Fuzzy C-means Clustering – An Approach to access Student's Performance in E-Learning Environment"*, International Journal of Advanced Research in Computer Science, Vol. 9, No. 1, pp. 887-892, 2018.

[13] Weina Wang and Yunije Zhang, *"On Fuzzy Cluster Validity Indices"*, Fuzzy Sets and Systems, Vol. 158, Issue 19, pp. 2095-2117, 2007.

Authors Profile

Dr. O.A. Mohamed Jafar obtained his M.Phil. in Computer Science from Bharathidasan University, Tiruchirappalli, Tamil Nadu, India in 1998. He has completed Ph.D. in Computer Science from Bharathidasan University, Tiruchirappalli, Tamil Nadu, India in 2015. He is currently working as Associate Professor of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has published 15 research papers in reputed National and International Journals. His areas of interest includes Data Mining, Clustering, Swarm Intelligence and Evolutionary Algorithms. He has 29 years of teaching experience.