Fuzzy Linear Discriminant Analysis Clustering With Its Application

Iden .H. Alkanani[1] and Rand .M. Fawzi[2*]

[1] Department of mathematics, College of Science for Women , University of Baghdad, Baghdad, Iraq.
[2] Department of mathematics , College of Education for Pure Science , Ibn-Al-Haitham, University of Baghdad, Baghdad, Iraq.

Abstract

Many fuzzy clustering are based on within-cluster scatter with a compactness measure , but in this paper explaining new fuzzy clustering method which depend on within-cluster scatter with a compactness measure and between-cluster scatter with a separation measure called the fuzzy compactness and separation (FCS). The fuzzy linear discriminant analysis (FLDA) based on within-cluster scatter matrix and between-cluster scatter matrix . Then two fuzzy scattering matrices in the objective function assure the compactness between data elements and cluster centers .To test the optimal number of clusters using validation clustering method is discuss .After that an illustrate example are applied.

Keywords: clustering , fuzzy compactness and separation (FCS), fuzzy linear discriminant analysis (FLDA), validation clustering method .

التحليل المميز الخطي للعنقدة الضبابية مع تطبيق

ايدن حسن الكناني[*1] و رند مهند فوزي[*2]

[1]قسم الرياضيات ، كلية العلوم للبنات ، جامعة بغداد ، بغداد ، العراق.

[2]قسم الرياضيات ، كلية التربية للعلوم الصرفة ابن الهيثم ، جامعة بغداد ،بغداد ، العراق.

الخلاصة

العديد من طرائق العنقدة تعتمد على انتشار العناصر داخل العناقيد بأستخدام مقياس التراص ، ولكن في هذا البحث سوف نشرح طريقة العنقدة الضبابية والتي سوف تعتمد على انتشار العناصر داخل العناقيد بأستخدام مقياس التراص بالأضافة الى انتشار العناصر بين العناقيد بأستخدام مقياس الأنفصال والتي تدعى بعنقدة ألتراص وألأنفصال الضبابية.

ان التحليل المميز الخطي الضبابي يعتمد على مصفوفة انتشار العناصر داخل العناقيد و مصفوفة انتشار العناصر بين العناقيد ، لذا فأن مصفوفتي ألأنتشار الضبابية نعتمد عليها في ايجاد دالة الهدف لهذه الطريقة.

ولأختيار العدد الأمثل من العناقيد فأننا سوف نستخدم طريقة صحة العنقدة الضبابية ومن ثم نطبق مثال توضيحي على العنقدة الضبابية.

* Email: rand_moh88@yahoo.com

Introduction

Cluster analysis is a branch in statistical multivariate analysis and unsupervised pattern recognition learning . The main objective of clustering is to classified set of elements data to cluster , where the elements data in cluster is more similar for each other and dissimilar for different cluster , that means , it classified .Data set into most similar groups in the same cluster and most dissimilar groups in different clusters .

Clustering analysis is a tool that assesses the relationship among samples of data set by organizing patterns into different groups , such that patterns within one group (cluster) show greater similarity to each other than those belonging to different groups .

There are two kinds of clustering , the first one was called by hard clustering which introduced by Tryon in 1939, [1] , the idea of this kind is that each elements of data belong to one cluster only , that means $X_{it} \in \{0,1\}$. The second one was called by fuzzy clustering which introduced by Ruspini in 1970, [2] , the idea of this kind that each elements may belong to all clusters or some clusters with different membership , that mean $X_{it} \in [0,1]$.

The aim of hard and fuzzy clustering is to partitioned the data matrix which contains (n)observations and (P) patterns into number of cluster (K) , therefore collect the observations which similar and more closer in cluster .

Fuzzy clustering plays an important role in pattern recognition , image processing and data analysis . In fuzzy clustering every point is assigned a membership to represent the degree of belonging to a certain group (cluster).

This paper organize as follows : section two contains the concept of fuzzy compactness and separation . Section three contain the fuzzy linear discriminant analysis . Section four explain the methodelegy of fuzzy linear discriminant analysis . Section five contains clustering validation method . section six contains the result and conclusion an illustrate example with conclusions .

1- Fuzzy compactness and separation :

There are many fuzzy clustering methods which depend on Euclidean distance or any other distances . One of the most well-known cluster is fuzzy C-mean (FCM) method which introduced by Bezdek in 1980 [3].

The (FCM) method based only on the sum of distance between observations in the samples to their cluster centers , which is equal to the trace of the within-cluster scatter matrix . This method based on minimizing the within-cluster scatter matrix trace the within-cluster scatter matrix trace interpreted as a compactness measure with a within-cluster variation .

Ozdemir and Akarun in 2001, [4], proposed an inter-cluster separation clustering algorithm that involves a separation measure in the inter-cluster separation objective function , because the between-cluster scatter matrix trace can be interpreted as a separation measure with between-cluster variation , maximization of the between-cluster scatter matrix trace will induce a result with will-separated cluster .

Wu , Yu and Yang in 2005, [5] proposed a novel fuzzy clustering algorithm called the fuzzy compactness and separation (FCS) algorithm . The (FCS) objective function is based on fuzzy scatter matrix , which derived by minimizing the compactness measure simultaneously maximizing the separation measure .

The compactness is measured using a fuzzy within-cluster scatter matrix trace , while the separation measured using a fuzzy between-cluster scatter matrix trace . The FCS algorithm is more robust to noise and others than FCM algorithm when weighting exponent (m) is large .

2- Fuzzy linear discriminant analysis : in recent years , linear discriminant analysis (LDA) [5] is most popular technique for data classification and dimensionality reduction in supervised classification problems . The LDA make data classification and doesn't change the location of the original data sets but only tries to provide more class separation and draw decision region between the given classes .

LDA uses the mean vector and covariance matrix of each class to formulate within-class , between –class , and mixture-class matrices LDA method can be used to project high-dimensional data into a low dimensional space . The LDA method maximizing between-class scatter and minimizing within-class scatter simultaneously in the projective feature space .

By using the concept of class scattering to class separation , the fisher criterion takes the

large values from sample when they are well clustered a round their mean within each class and the cluster of the different classes are well separated.

The LDA optimizes class separation by maximize ratio of interclass to interclass variance and is suitable for applications with unequal sample sizes. The aim of this paper is to study and discuss the fuzzy linear discriminant analysis for clustering and apply it for illustrate example , by choosing many assumed clusters to show which number of clusters is optimal by minimizing the objective function and use the validation clustering method to test the similar of elements (point) in cluster and to know the optimal number of clusters.

3- Unsupervised Fuzzy Linear Discrimination Analysis Method :

The steps of the unsupervised fuzzy linear discrimination analysis (UFLDA) are :

1-Let you have the data matrix (X_{it}) with dimension (nxp ,i=1,....,n, t=1,…..,p) where n represent the samples or observations and P are attributes or features

2-We generate the partition matrix (U_{ij}) this matrix of dimension (nxk , i=1,….,n , j=1,….,k) where k represent to the clusters

This matrix generation randomly with uniform distribution

$$f(u) = \begin{cases} 1/(b-a) & a<u<b \\ 0 & o.w \end{cases}$$

a- Each element of the matrix U are random with $U_{ij} \in [0,1]$.

b-The sum of each row for this matrix is equal to the one $\sum_{j=1}^{k} u_{ij} = 1$

c- The sum of each column for this matrix must be less or equal to the samples size (n)

$0 \leq \sum_{i=1}^{n} u_{ij} \leq n$

d- Determining the C-means matrix (C_{jt}) with dimension (kxp, j=1,….,k, t=1,…..,p)which represent the center of clusters by the formula

$$C_{jt} = \sum_{j=1}^{k} \left(\frac{u_{ij}}{\sum_{i=1}^{n} u_{ij}} \right) x_{it} \quad where\ C_i\ is\ the\ class\ mean$$

And compute C represent the total mean by the formula

$C = \frac{1}{N} \sum_{i=1}^{n} x_j$ where N represent to all element

in data matrix

3-Find the between-cluster scatter matrix (S_b^{UFLDA}) by the formula :

$$S_b^{UFLDA} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} u_{ij}}{N} (C_i - C)(C_i - C)'$$

And also find the within-cluster scatter matrix (S_w^{UFLDA}) by the formula:

$$S_w^{UFLDA} = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{u_{ij}}{N} (x_j - C_i)(x_j - C_i)'$$

4-Finally ,computing the objective function by the formula :

$$J_{FLDC}(u_{ij}) = tr\left[(S_w^{UFLDA})^{-1} (S_b^{UFLDA})\right]$$

4- Clustering validation method:

To test the similarity between elements (points) in cluster to know the optimal number of clustering from unsupervised classification and to prove the validity of clustering , we must use the clustering validation method there is many methods for validity clustering .In this research we use (FS) Fukuyama and Sugeno validation is proposed in 1989, see [6] we can find it by the formula :

$$FS = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m} \left\| x_j - c_i \right\|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m} \left\| c_i - \bar{c} \right\|^2$$

$$Where\ \bar{c} = \sum_{i=1}^{n} c_i / n$$

In this function the first term is measure the compactness of the clusters and the second one measures of the distances of the clusters representatives ,it is clear for compact and well-separated clusters we expect small values for FS

5- Result and conclusion: in this section giving two illustrate examples , first we assume that the number of clusters is (3) with data set matrix ,second we assume that the number of clusters is (4) with the same data set matrix , then compare between these two examples to know the optimal value of clustering.

Example (1): taking the data matrix X of dimension (16 x 4) and partition matrix U of dimension (16 x 3):

$$X=\begin{bmatrix}2&0&1&4\\3&2&2&1\\2&1&0&1\\3&5&6&7\\4&1&5&7\\2&4&2&8\\1&4&1&0\\6&2&1&5\\1&5&1&0\\0&7&1&3\\5&1&2&5\\6&0&8&4\\1&5&1&0\\7&3&5&9\\0&1&1&5\\5&1&3&2\end{bmatrix}, U=\begin{bmatrix}0.2&0.6&0.2\\0.2&0.1&0.7\\0.4&0.1&0.5\\0.3&0.3&0.4\\0.3&0.6&0.1\\0.8&0.1&0.1\\0.1&0.7&0.2\\0.3&0.6&0.1\\0.4&0.5&0.1\\0.1&0.5&0.4\\0.4&0.5&0.1\\0.5&0.4&0.1\\0.3&0.6&0.1\\0.8&0.1&0.1\\0.8&0.1&0.1\\0.5&0.4&0.1\end{bmatrix}$$

$$C=\begin{bmatrix}3.3125&2.3594&2.7813&4.6875\\2.9355&2.7581&2.3710&3.3065\\2.5294&2.8824&2.2059&3.0882\end{bmatrix}$$

$$SB=\begin{bmatrix}0.0184&0.0207&0.0289&0.0238\\0.0224&0.0451&0.0138&0.0224\\0.0270&0.0148&0.0270&0.0315\\0.0224&0.0451&0.0451&0.0315\end{bmatrix}$$

$$SW=\begin{bmatrix}1.1656&-0.5157&0.7174&0.6858\\-0.5157&1.0463&-0.2056&-0.1654\\0.7174&-0.2056&1.2046&0.7781\\0.6858&-0.1654&0.7781&1.9401\end{bmatrix}$$

J = 0.092
FS = 27.4040

Example (2): taking data matrix X of dimension (16 x 4) and partition matrix U of dimension (16 x 4):

$$X=\begin{bmatrix}2&0&1&4\\3&2&2&1\\2&1&0&1\\3&5&6&7\\4&1&5&7\\2&4&2&8\\1&4&1&0\\6&2&1&5\\1&5&1&0\\0&7&1&3\\5&1&2&5\\6&0&8&4\\1&5&1&0\\7&3&5&9\\0&1&1&5\\5&1&3&2\end{bmatrix}, U=\begin{bmatrix}0.2&0.6&0.1&0.1\\0.2&0.1&0.5&0.2\\0.2&0.2&0.1&0.5\\0.3&0.3&0.3&0.1\\0.3&0.5&0.1&0.1\\0.5&0.3&0.1&0.1\\0.1&0.4&0.3&0.2\\0.2&0.1&0.6&0.1\\0.4&0.3&0.2&0.1\\0.1&0.5&0.1&0.3\\0.1&0.3&0.5&0.1\\0.2&0.4&0.1&0.3\\0.2&0.1&0.6&0.1\\0.7&0.1&0.1&0.1\\0.5&0.2&0.2&0.1\\0.3&0.2&0.4&0.1\end{bmatrix}$$

$$C=\begin{bmatrix}3.2222&2.6444&2.8000&4.7556\\2.6957&2.5870&2.6522&3.9348\\3.2326&2.6977&2.1163&3.1395\\2.7692&2.5385&2.3462&3.0769\end{bmatrix}$$

$$SB=\begin{bmatrix}0.0247&0.0217&0.0210&0.0269\\0.0279&0.0345&0.0195&0.0221\\0.0307&0.0196&0.0193&0.0228\\0.0221&0.0410&0.0340&0.0264\end{bmatrix}$$

$$SW=\begin{bmatrix}1.1716&-0.5343&0.7376&0.7259\\-0.5343&1.0578&-0.2174&-0.2053\\0.7376&-0.2174&1.1999&0.7771\\0.7259&-0.2053&0.7771&1.9533\end{bmatrix}$$

J = 0.0888
FS = 20.1727

Conclusion: from the result of J and FS we can make the following conclusion :

1-When comparing between the objective function (J) for three and four clusters, we find that the smaller objective function (J) if the number of cluster is four
2-When comparing between the FS validity method for three and four clusters, we find that the smaller FS validity if the number of clusters is four.
3-Concluding that the optimal number of clusters for the study data is four

Samples	**Cluster**	**Objective function**	**FS**
16	3	0.9601	27.404
16	4	0.092	20.1727

.

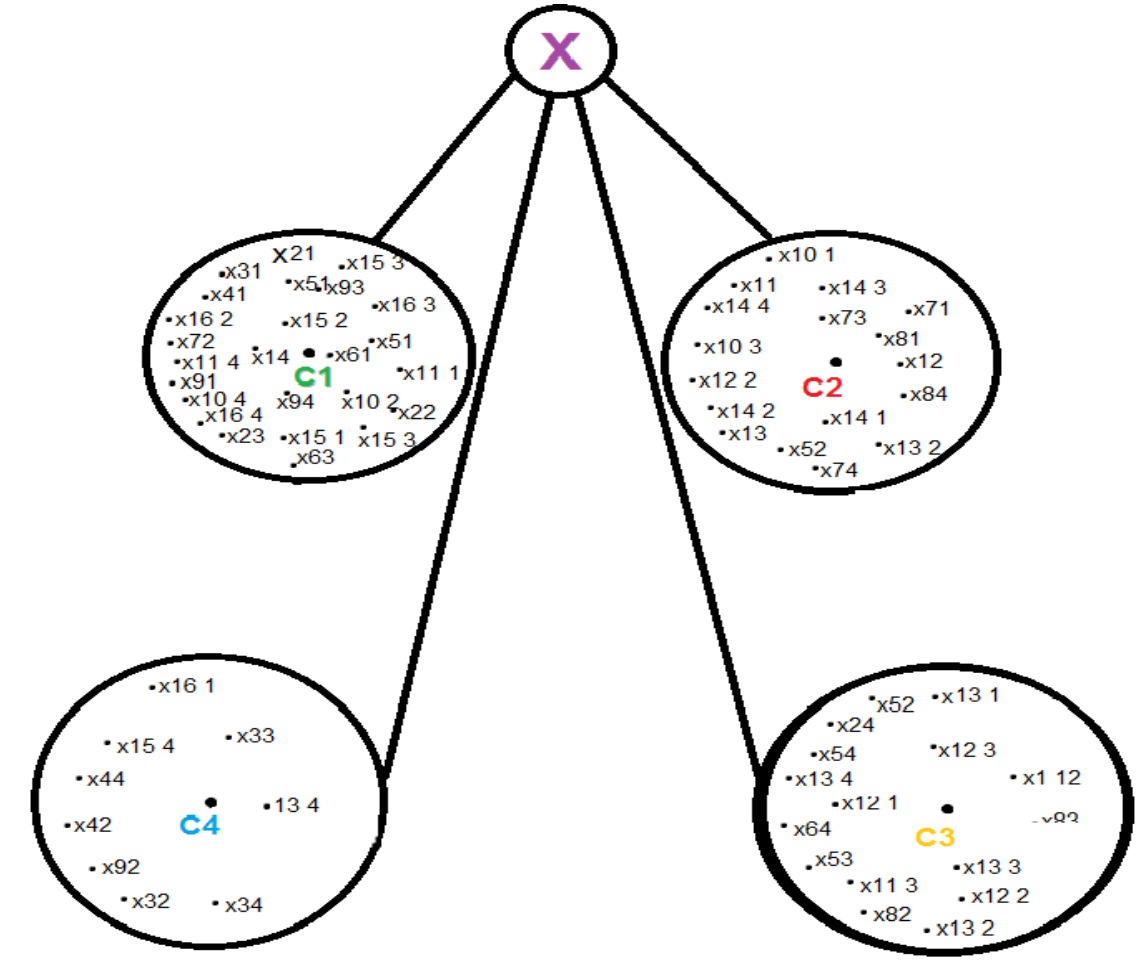

Figure 1- Represent the distributed of elements (X) on clusters ,From this figure , we distributed the elements (Xit) which means the data matrix on the cluster

Reference

1.Goodgman, L. and Kruskal , w.,**1954**" Measures of associations for cross-validations ", JASA , Vol 49 ,P 732-764.

2.Ruspini , E.H. , **1970**, "Numerical methods for fuzzy clustering " , Information science , Vol 2 , P 319-350 .

3.Bezdek ,J.C. **1980**.,"A convergence theorem for the fuzzy ISODATA clustering algorithms ", IEEE ,Vol2 , P 1-8

4.Ozdemir D. , and Akarun L. , **2001**,"fuzzy algorithms for combined quantization and dithering ", IEEE Trans. Image processing, Vol 6(10) , P923-931.

5.Wu K. , Yu J. and Yang M. , **2005**,"A novel fuzzy clustering algorithm based on fuzzy scatter matrix with optimality tests ", Pattern recognition letters , Vol 26 , P 639-652 .

6.Halkidi,M. ,Batistakis,Y. and Vazirgiannis,M. ,**2001**,"On clustering validation techniques ".Journal of intelligent inf. System ,17:2/3,pp:107-145.