Pattern Recognition using the Fuzzy c-means Technique

Samarjit Das

Department of Computer Science & IT, Cotton College, Guwahati-01, Assam, India
ssaimm@rediffmail.com

Abstract

In the field of pattern recognition due to the fundamental involvement of human perception and inadequacy of standard Mathematics to deal with its complex and ambiguously defined system, different fuzzy techniques have been applied as an appropriate alternative. A pattern recognition system has to undergo basically the steps of preprocessing, feature extraction and selection, classifier design and optimization. In our work the data we have analyzed is in the form of numerical vectors, with a number of clusters predefined. Therefore the fuzzy c-means technique of Bezdek has been considered for our work. Although in the fuzzy c-means technique Euclidean distance has been used to obtain the membership values of the objects in different clusters, in our present work along with Euclidean distance we have used other distances like Canberra distance, Hamming distance to see the differences in outputs.

Keywords: *Pattern recognition, fuzzy c-means technique, Euclidean distance, Canberra distance, Hamming distance*

1. Introduction

The use of fuzzy set theory (FST), developed by Zadeh [1], has proliferated the research work especially in the field of modeling uncertainty. A complete presentation of all aspects of FST is available in the work of Zimmermann [2]. The applications of FST in dealing with ambiguous problems where uncertainty prevails have been reflected in the works of Dewit [3], Lemaire [4], Ostaszewski [5]. Pattern recognition is a field whose objective is to assign an object or event to one of a number of categories, based on features derived to emphasize commonalities. Zheng and He [6] reviewed the general processing steps of pattern recognition where they have discussed several methods used for the steps of pattern recognition such as Principal Component Analysis (PCA) in feature extraction, Support Vector Machines (SVM) in classification and so forth. Derring and Ostaszewski [7] have explained in their research work a method of pattern recognition for risk and claim classification. They have also made similar application to classify claims with regard to their suspected fraud content. Bezdek [8] has discussed in his fuzzy c-means technique that the data to be analyzed must be in the form of numerical vectors called feature vectors, and the number of clusters must be predefined for obtaining the membership values of the feature vectors. Bezdek and Pal [9] have described it in their classification technique as a numerical process description, the fuzzy c-means iterative algorithm.

Sir Francis Galton [10] was the first psychologist who devoted his time in the study of individual differences. Each individual exhibits certain characteristics or features due to which it is possible to measure the degree of similarity between two individuals or to notice a minimal difference between two individuals. In our present study each feature vector (i.e. object or student) consists of three features namely Intelligent Quotient (IQ), Achievement Motivation (AM) and Social Adjustment (SA). We have predefined five cluster namely very superior (C1), high average (C2), average (C3), low average (C4) and borderline defective

(C5). Using fuzzy classification technique we will see that depending on degree of membership of each object in different clusters how the objects partially or fully belong to clusters. The respective ranges of initial clusters against each feature have been shown in Table 2 of Section 3. The data we have analyzed for students' individual differences have been collected from different schools of Guwahati, Assam, India. The raw scores of the data for our present work have been given in table1 of Section 3. In our work the dataset is in the form of numerical vectors, with five clusters and three features predefined. Therefore the fuzzy c-means technique of Bezdek [8] has been considered for our work. Although in the fuzzy c-means technique Euclidean distance has been used to obtain the membership values of the objects in different clusters, in our present work along with Euclidean distance we have used other distances like Canberra distance, Hamming distance to see the differences in outputs.

In Section 2 we have discussed the concept of fuzzy set and the mathematical algorithms needed to implement classification using fuzzy techniques. The findings of our work have been placed in Section 3. In Section 4 we make a comparison of the findings with respect to different distances. Section 5 consists of the field of application. The conclusion is given in Section 6.

2. Fuzzy Set and Mathematical Algorithm for Classification

A fuzzy subset $\tilde{A}$ of X, universe of discourse, is defined by its membership function $\mu_{\tilde{A}}: X \rightarrow [0, 1]$. For any $x \in X$, the value $\mu_{\tilde{A}}(x)$ specifies the degree to what x belongs to $\tilde{A}$. As uncertainty prevails in the field of finding individual differences in our work, therefore a fuzzy technique is required for the classification of the same. In our work we use a fuzzy pattern recognition technique given by Bezdek [8]. The basic task of a classification technique is to divide n patterns, where n is a natural number , represented by vectors in a p-dimensional Euclidean space, into c, $2 \leq c < n$, categorically homogeneous subsets which are called clusters. Let the data set be

$X = \{x_1, x_2, \ldots\ldots\ldots, x_n\}$, where $x_k = \{x_{k1}, x_{k2}, \ldots\ldots\ldots, x_{kp}\}$, k= 1,2,3,……..,n.

Each x_k is called a feature vector and x_{kj} where j=1,2,…..p is the j^{th} feature of the k^{th} feature vector.

A partition of the dataset X into clusters is described by the membership functions of the elements of the cluster. Let S_1, S_2,……..,S_c denote the clusters with corresponding membership functions μ_{S_1} , μ_{S_2} , ……., μ_{S_c}

A c x n matrix containing the membership values of the objects in the clusters

$\tilde{U} = [\mu_{S_i}(x_k)]_{i=1,2,\ldots\ldots c, k=1,2,\ldots\ldots n}$ is a fuzzy c- partition if it satisfies the following conditions

$$\sum_{i=1}^{c} \mu_{s_i}(x_k) = 1 \qquad \text{for each k= 1,2,….,n.} \tag{1}$$

$$0 \leq \sum_{k=1}^{n} \mu_{s_i}(x_k) \leq n \quad \text{for each i= 1,2,….,c.} \tag{2}$$

Condition (1) says that each feature vector x_k has its total membership value 1 divided among all clusters. Condition (2) states that the sum of membership degrees of feature vectors

in a given cluster does not exceed the total number of feature vectors. The algorithm of the fuzzy c-means technique of Bezdek [8] has been illustrated below.

Step 1: Choose the number of clusters, c, 2≤c<n, where n is the total number of feature vectors. Choose m, 1≤ m <α. Define the vector norm ‖ ‖(generally defined by the Euclidean distance), *i.e.*,

$$\| x_k - v_i \| = \sqrt{\sum_{j=1}^{p} (x_{kj} - v_{ij})^2} \qquad (3)$$

where x_{kj} is the j[th] feature of the k[th] feature vector, for k=1,2,……,n; j=1,2,….,p and v_{ij}, j-dimensional centre of the i[th] cluster for i=1,2,……,c; j=1,2,….,p; n, p and c denote the total number of feature vector , features in each feature vector and total number of clusters respectively.

Choose the initial fuzzy partition

$$U^{(0)} = [\mu_{s_i}{}^{(0)}(x_k)]_{1\le i\le c, 1\le k\le n}$$

Choose a parameter ∈>0 (this will tell us when to stop the iteration). Set the iteration counting parameter l equal to 0.

Step 2: Calculate the fuzzy cluster centers $\{v_i^{(l)}\}_{i=1,2,.....c}$ given by the following formula

$$v_i^{(l)} = \frac{\sum_{k=1}^{n} (\mu_{s_i}{}^{(l)}(x_k))^m x_k}{\sum_{k=1}^{n} (\mu_{s_i}{}^{(l)}(x_k))^m} \qquad \text{for i = 1, 2 , ….. c;} \quad \text{k= 1, 2, , …..n.} \qquad (4)$$

Step 3: Calculate the new partition matrix (i.e. membership matrix)

$$U^{(l+1)} = [\mu_{s_i}{}^{(l+1)}(x_k)]_{1\le i\le c, 1\le k\le n} \quad \text{, where } \mu_{s_i}{}^{(l+1)}(x_k) = \frac{1}{\sum_{j=1}^{c} \left(\frac{\| x_k - v_i^{(l)} \|}{\| x_k - v_j^{(l)} \|}\right)^{\frac{2}{m-1}}}$$

(5)

for i=1,2,……..,c and k=1,2,……..,n. If $x_k = v_i^{(l)}$, formula (5) cannot be used. In this case the membership function is

$$\mu_{s_i}{}^{(l+1)}(x_k) = \begin{cases} 1 if k=i \\ 0 if k\ne i, i=1,2,....c \end{cases}$$

(6)

Step 4: Calculate $\Delta = \| U^{(l+1)} - U^{(l)} \|$

If $\Delta > \in$, repeat steps 2, 3 and 4. Otherwise, stop at some iteration count l^*. To make the result operational the fifth step had been introduced by Derring and Ostaszewski [7].

Step 5: The final fuzzy matrix U^{l^*} is structured for operational use by means of the normalized α -cut, for some 0 < α< 1. All membership values less than α are replaced with zero and the function is renormalized (sums to one) to preserve partition condition (1).

3. Our Work and Findings

In our present work the dataset we analyze consists of fifty (50) feature vectors with three (03) features (Intelligent Quotient (IQ), Achievement Motivation (AM) and Social Adjustment (SA)) each. We have predefined five (05) clusters namely very superior (C1), high average (C2), average (C3), low average (C4) and borderline defective (C5). The respective ranges of initial clusters against each feature have been shown in Table 2. The raw scores of the data for our present work have been given in Table 1. The fuzzy c-means technique of Bezdek [8] (see the algorithm in Section 2) has been used in our work as it has been considered to be the most appropriate. The membership values in initial partitions have been calculated by step1 of the algorithm (see Section 2). For this purpose, only one feature, IQ, out of the three features has been considered, *i.e.*, depending on the value of the feature, IQ, of the feature vector, it will belong to any one of the clusters. Therefore initially the membership value of a feature vector to a cluster will be either 0 or 1. Table 3 shows the membership values of the feature vectors in initial fuzzy partitions.

Table 1. Scores of Students Against Three Features

OBJ	IQ	AM	SA	OBJ	IQ	AM	SA	OBJ	IQ	AM	SA
1	91	18	55	18	130	19	75	35	125	19	85
2	85	16	40	19	90	17	55	36	80	18	60
3	120	19	74	20	91	17	56	37	85	18	70
4	90	18	75	21	140	22	82	38	145	25	90
5	92	17	74	22	92	18	75	39	80	18	74
6	82	17	55	23	101	18	55	40	92	17	55
7	95	19	75	24	85	16	54	41	120	18	70
8	89	18	74	25	97	19	54	42	145	30	80
9	96	19	75	26	110	18	55	43	95	18	50
10	90	17	55	27	100	16	40	44	80	16	36
11	97	16	54	28	100	18	75	45	90	17	55
12	125	21	74	29	70	14	30	46	115	23	84
13	100	19	75	30	105	17	55	47	100	18	80
14	90	17	54	31	79	14	35	48	80	14	35
15	100	18	84	32	80	15	34	49	105	19	75
16	95	19	75	33	125	20	75	50	120	21	74
17	130	23	85	34	100	19	75				
OBJ: ID of student IQ: Intelligent Quotient , AM: Achievement motivation, SA: Social adjustment											

Table 2. Initial Cluster of Individual Difference

	IQ	AM	SA
Borderline defective	70-79	11-14	30-34
Low average	80-89	15-16	35-54
Average	90-109	17-18	55-74
High average	110-139	19-22	75-84
Very superior	140-169	23 & above	85 & above
IQ: Intelligent Quotient , AM: Achievement motivation, SA: Social adjustment			

Table 3. Membership Values of the Feature Vectors in the Initial Fuzzy Partition

FUZZY MEMBERSHIP VALUES

OBJ	C1	C2	C3	C4	C5	SUM
1	0	0	1	0	0	1
2	0	0	0	1	0	1
3	0	1	0	0	0	1
4	0	0	1	0	0	1
5	0	0	1	0	0	1
6	0	0	0	1	0	1
7	0	0	1	0	0	1
8	0	0	0	1	0	1
9	0	0	1	0	0	1
10	0	0	1	0	0	1
11	0	0	1	0	0	1
12	0	1	0	0	0	1
13	0	0	1	0	0	1
14	0	0	1	0	0	1
15	0	0	1	0	0	1
16	0	0	1	0	0	1
17	0	1	0	0	0	1
18	0	1	0	0	0	1
19	0	0	1	0	0	1
20	0	0	1	0	0	1
21	1	0	0	0	0	1
22	0	0	1	0	0	1
23	0	0	1	0	0	1
24	0	0	0	1	0	1
25	0	0	1	0	0	1
26	0	1	0	0	0	1
27	0	0	1	0	0	1
28	0	0	1	0	0	1
29	0	0	0	0	1	1
30	0	0	1	0	0	1
31	0	0	0	0	1	1
32	0	0	0	1	0	1
33	0	1	0	0	0	1
34	0	0	1	0	0	1
35	0	1	0	0	0	1
36	0	0	0	1	0	1
37	0	0	0	1	0	1
38	1	0	0	0	0	1
39	0	0	0	1	0	1
40	0	0	1	0	0	1
41	0	1	0	0	0	1
42	1	0	0	0	0	1
43	0	0	1	0	0	1
44	0	0	0	1	0	1
45	0	0	1	0	0	1
46	0	1	0	0	0	1
47	0	0	1	0	0	1
48	0	0	0	1	0	1
49	0	0	1	0	0	1
50	0	1	0	0	0	1
SUM	3	10	25	10	2	

C-mean Fuzzy Clustering with 3 coverage data pattern after 0 iteration ,where epsilon =0.05

After finding the membership values of the feature vectors in the initial partitions we combine all three features (IQ, AM and SA) (see step 3 of algorithm of Section 2) and find the membership values of the feature vectors in the next level of partition. This process continues till the value of $\Delta > \in$. In each case the values of the variables while implementing the algorithm have been considered as given below

n=50, p=3, c=5, m=2, epsilon ($\in$) = 0.05, aCut (α-cut) = 0.20.

In our present work along with Euclidean distance we have tested two more distances namely Hamming distance and Canberra distance to find the distances of the feature vectors from the cluster centers of different clusters and based on which the membership values of the feature vectors in different clusters have been calculated.

Let $x_k = \{x_{k1}, x_{k2}, ..., x_{kp}\}, v_i = \{v_{i1}, v_{i2}, ..., v_{ip}\}$ be two p-dimensional vectors.

The vector norms $\| x_k - v_i \|$ with respect to the three different distances used in our work have been given in the following

Euclidean distance $$\| x_k - v_i \| = \sqrt{\sum_{j=1}^{p} (x_{kj} - v_{ij})^2} \quad (7)$$

Canberra distance $$\| x_k - v_i \| = \sum_{j=1}^{p} \frac{| x_{kj} - v_{ij} |}{| x_{kj} | + | v_{ij} |} \quad (8)$$

Hamming (or city block) distance $$\| x_k - v_i \| = \sum_{j=1}^{p} | x_{kj} - v_{ij} | \quad (9)$$

First we have tested our dataset by defining the vector norm by Euclidean Distance (see equation 7) in the algorithm (see Section 2). We have reached the final clusters after 4th iteration. The membership values of the feature vectors in the final clusters after taking α – cut =0.20 have been shown in Table 4. A graphical representation of the results in Table 4 has been shown in Figure 1. Here full membership value and partial membership value of a feature vector in a cluster have been represented by a diamond and a rectangle respectively.

Next we have used the same algorithm (see Section 2) on the same dataset (see Table 1) but by defining the vector norm by Canberra distance (see equation 8) to see the differences of membership values of the feature vectors in different clusters. The initial partitions remain same as shown in Table 3. With this vector norm we have reached the final clusters after 13th iterations. The membership values of the feature vectors in the final clusters after taking α – cut =0.20 have been shown in Table 5. Figure 2 represents the graphical view of the results in Table 5.

After Canberra distance we have tested the same dataset (see Table 1) with the same algorithm (see Section 2) by defining the vector norm by Hamming distance, also called city block distance (see equation 9). Here also the initial partitions remain same as shown in Table 3. With this vector norm we have reached the final clusters after 8th iterations. The membership values of the feature vectors in the final clusters after taking α – cut =0.20 have been shown in Table 6. A graphical representation of the results in Table 6 has been shown in Figure 3.

Table 4. Membership Values of the Feature Vectors in the Final Clusters after taking α – cut =0.20 while Defining the Vector Norm by Euclidean Distance

FUZZY MEMBERSHIP VALUES

OBJ	C1	C2	C3	C4	C5	SUM
1	0	0	0	1	0	1
2	0	0	0	0	1	1
3	0	1	0	0	0	1
4	0	0	1	0	0	1
5	0	0	1	0	0	1
6	0	0	0	1	0	1
7	0	0	1	0	0	1
8	0	0	1	0	0	1
9	0	0	1	0	0	1
10	0	0	0	1	0	1
11	0	0	0	1	0	1
12	0	1	0	0	0	1
13	0	0	1	0	0	1
14	0	0	0	1	0	1
15	0	0	1	0	0	1
16	0	0	1	0	0	1
17	0.5742	0.4258	0	0	0	1
18	0.2445	0.7555	0	0	0	1
19	0	0	0	1	0	1
20	0	0	0	1	0	1
21	1	0	0	0	0	1
22	0	0	1	0	0	1
23	0	0	0	1	0	1
24	0	0	0	1	0	1
25	0	0	0	1	0	1
26	0	0.2964	0.2727	0.4309	0	1
27	0	0	0	0.5732	0.4268	1
28	0	0	1	0	0	1
29	0	0	0	0	1	1
30	0	0	0.3408	0.6592	0	1
31	0	0	0	0	1	1
32	0	0	0	0	1	1
33	0	1	0	0	0	1
34	0	0	1	0	0	1
35	0.2982	0.7018	0	0	0	1
36	0	0	0	1	0	1
37	0	0	0.6543	0.3457	0	1
38	1	0	0	0	0	1
39	0	0	0.6462	0.3538	0	1
40	0	0	0	1	0	1
41	0	1	0	0	0	1
42	1	0	0	0	0	1
43	0	0	0	1	0	1
44	0	0	0	0	1	1
45	0	0	0	1	0	1
46	0	1	0	0	0	1
47	0	0	1	0	0	1
48	0	0	0	0	1	1
49	0	0.2925	0.7075	0	0	1
50	0	1	0	0	0	1
SUM	4.1169	8.472	14.6216	16.3627	6.4268	

C-mean Fuzzy Clustering with 3 coverage data pattern after 4 iteration ,where epsilon =0.05

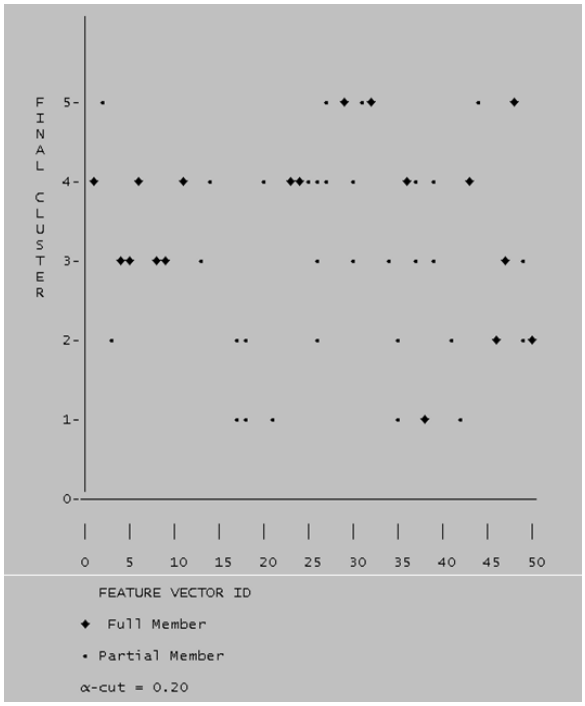

Figure 1. Partial and Full Membership of Feature Vectors in Final Clusters after Taking α – cut =0.20 while Defining the Vector Norm by Euclidean Distance

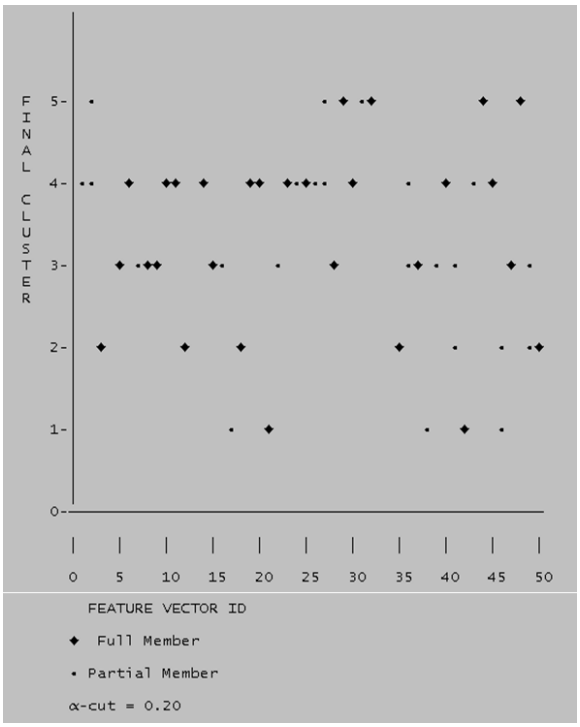

Figure 2. Partial and Full Membership of Feature Vectors in Final Clusters after Taking α – cut =0.20 while Defining the Vector Norm by Canberra Distance

Table 5. Membership Values of the Feature Vectors in the Final Clusters after Taking α – cut =0.20 while Defining the Vector Norm by Canberra Distance

FUZZY MEMBERSHIP VALUES

OBJ	C1	C2	C3	C4	C5	SUM
1	0	0	0	1	0	1
2	0	0	0	0.2916	0.7084	1
3	0	1	0	0	0	1
4	0	0	1	0	0	1
5	0	0	1	0	0	1
6	0	0	0	1	0	1
7	0	0	1	0	0	1
8	0	0	1	0	0	1
9	0	0	1	0	0	1
10	0	0	0	1	0	1
11	0	0	0	1	0	1
12	0	1	0	0	0	1
13	0	0	1	0	0	1
14	0	0	0	1	0	1
15	0	0	1	0	0	1
16	0	0	1	0	0	1
17	1	0	0	0	0	1
18	0	1	0	0	0	1
19	0	0	0	1	0	1
20	0	0	0	1	0	1
21	1	0	0	0	0	1
22	0	0	1	0	0	1
23	0	0	0	1	0	1
24	0	0	0	1	0	1
25	0	0	0	1	0	1
26	0	0	0	1	0	1
27	0	0	0	0.4841	0.5159	1
28	0	0	1	0	0	1
29	0	0	0	0	1	1
30	0	0	0	1	0	1
31	0	0	0	0	1	1
32	0	0	0	0	1	1
33	0	1	0	0	0	1
34	0	0	1	0	0	1
35	0	1	0	0	0	1
36	0	0	0.357	0.643	0	1
37	0	0	1	0	0	1
38	1	0	0	0	0	1
39	0	0	1	0	0	1
40	0	0	0	1	0	1
41	0	0.7019	0.2981	0	0	1
42	1	0	0	0	0	1
43	0	0	0	1	0	1
44	0	0	0	0	1	1
45	0	0	0	1	0	1
46	0.6205	0.3795	0	0	0	1
47	0	0	1	0	0	1
48	0	0	0	0	1	1
49	0	0.3195	0.6805	0	0	1
50	0	1	0	0	0	1
SUM	4.6205	7.4009	15.3356	16.4187	6.2243	

C-mean Fuzzy Clustering with 3 coverage data pattern after 13 iteration ,where epsilon =0.05

Table 6. Membership Values of the Feature Vectors in the Final Clusters after Taking α – cut =0.20 while Defining the Vector Norm by Hamming Distance

after taking α-cut =0.2

FUZZY MEMBERSHIP VALUES

OBJ	C1	C2	C3	C4	C5	SUM
1	0	0	0	1	0	1
2	0	0	0	0	1	1
3	0	1	0	0	0	1
4	0	0	1	0	0	1
5	0	0	1	0	0	1
6	0	0	0	1	0	1
7	0	0	1	0	0	1
8	0	0	1	0	0	1
9	0	0	1	0	0	1
10	0	0	0	1	0	1
11	0	0	0	1	0	1
12	0	1	0	0	0	1
13	0	0	1	0	0	1
14	0	0	0	1	0	1
15	0	0	1	0	0	1
16	0	0	1	0	0	1
17	0.727	0.273	0	0	0	1
18	0	1	0	0	0	1
19	0	0	0	1	0	1
20	0	0	0	1	0	1
21	1	0	0	0	0	1
22	0	0	1	0	0	1
23	0	0	0	1	0	1
24	0	0	0	1	0	1
25	0	0	0	1	0	1
26	0	0	0	1	0	1
27	0	0	0	0.5432	0.4568	1
28	0	0	1	0	0	1
29	0	0	0	0	1	1
30	0	0	0	1	0	1
31	0	0	0	0	1	1
32	0	0	0	0	1	1
33	0	1	0	0	0	1
34	0	0	1	0	0	1
35	0.3054	0.6946	0	0	0	1
36	0	0	0	0.6403	0.3597	1
37	0	0	0.6631	0.3369	0	1
38	1	0	0	0	0	1
39	0	0	1	0	0	1
40	0	0	0	1	0	1
41	0	1	0	0	0	1
42	1	0	0	0	0	1
43	0	0	0	1	0	1
44	0	0	0	0	1	1
45	0	0	0	1	0	1
46	0.3947	0.6053	0	0	0	1
47	0	0	1	0	0	1
48	0	0	0	0	1	1
49	0	0.2723	0.7277	0	0	1
50	0	1	0	0	0	1
SUM	4.4271	7.8452	14.3907	16.5205	6.8165	

C-mean Fuzzy Clustering with 3 coverage data pattern after 8 iteration ,where epsilon =0.05

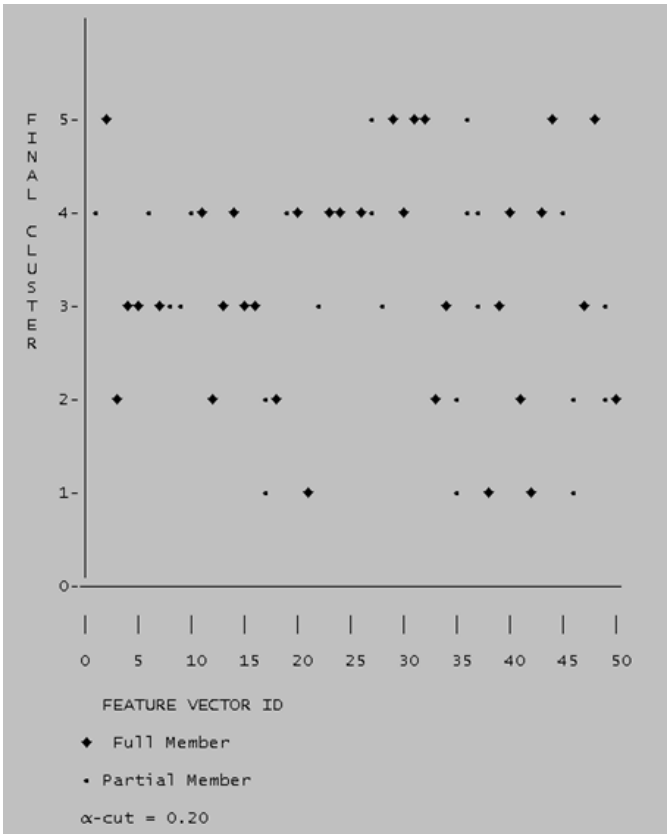

Figure 3. Partial and Full Membership of Feature Vectors in Final Clusters after Taking α – cut =0.20 while Defining the Vector norm by Hamming Distance

In our work C ++ has been used as a programming language to implement the fuzzy c-means algorithm. We have also used MS-Excel in the analysis section.

4. Analysis

In our work the vector norm has been defined with three different distances separately on the same algorithm and has been applied on the same dataset. It reveals the following differences in the results.

First, the algorithm takes the least number of iterations, only four, while the vector norm has been defined by Euclidean distance, to reach the final clusters (see Table 7). On the other hand the algorithm takes the most number of iterations, thirteen, while the vector norm has been defined by Canberra distance, to reach the final clusters (see Table 7). This implies that the algorithm produces the result fastest when Euclidean distance is considered to define the vector norm, and it produces the result slowest when the vector norm is defined by Canberra distance. The graphical representation of the same has been shown in Figure 4.

Secondly, we obtain cluster-wise total number of full membership and partial membership values of feature vectors with respect to three different vector norms defined by three different distances (see Table 8). The graphical representation of the same has been given in Figure 5. Here we observe that in most of the cases the least number of feature vectors exhibit

full membership values and the most number of feature vectors exhibit partial membership values in the final clusters while considering Euclidean distance to define the vector norm in the algorithm. On the other hand while considering Canberra distance to define the vector norm in the algorithm, in most of the cases the most number of feature vectors exhibit full membership values and the least number of feature vectors exhibit partial membership values in the final clusters.

Since individual difference in our dataset is not a crisp concept but a fuzzy one, therefore for each feature vector, instead of full membership to a single cluster, partial membership to more than one cluster is expected.

From this point of view we may say that out of the three distances, the algorithm shows the most expected result when the vector norm has been defined by Euclidean distance and the least expected one when the vector norm has been defined by Canberra distance.

Table 7. No. of Iterations after which the Loop Stops with Respect to Three Different Distances

EUCLIDEAN DISTANCE	4
CANBERRA DISTANCE	13
HAMMING DISTANCE	8
α-cut = 0.20 , $\in$= 0.05	

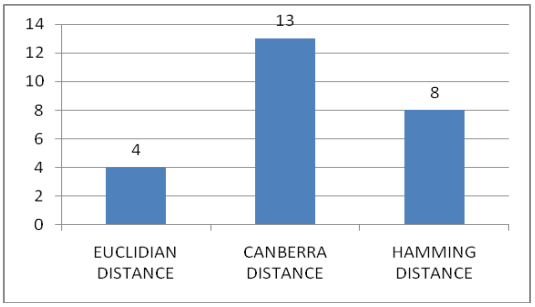

Figure 4. No. of Iterations after which the Loop Stops with Respect to Three Different Distances

Table 8. Cluster-wise Total No. of Full Membership and Partial Membership with Respect to Three Different Distances

	Full membership					Partial membership				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
EUCLIDEAN DISTANCE	3	6	12	14	6	3	5	5	5	1
CANBERRA DISTANCE	4	6	14	15	5	1	3	3	3	2
HAMMING DISTANCE	3	6	13	15	6	3	4	2	3	2
α- cut = 0.20 , $\in$= 0.05										

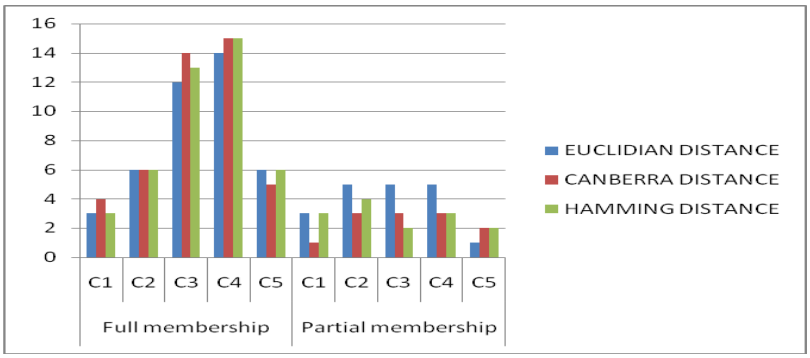

Figure 5. Cluster-wise Total No. of Full Membership and Partial Membership Values of Feature Vectors in Different Clusters with Respect to Three Different Distances

5. Application

We have tested the fuzzy c-means technique of Bezdek [8] (see the algorithm in Section 2) on a dataset of individual differences (see Table 1) to see the membership value of an individual in the clusters. As the data we have analyzed is in the form of numerical vectors, therefore the same algorithm can be applied in other relevant field like vehicular pollution (as for example) where the data is in the form of numerical vectors.

6. Conclusion

As the dataset we have analyzed is in the form of numerical vectors and the number of clusters has been predefined, the fuzzy c-means algorithm of Bezdek has been considered for the classification of the same. Although in general, Euclidean distance has been used in the fuzzy c-means algorithm, we tried it with two more distances namely Canberra distance and Hamming distance to see the differences in the results. It has been reflected in the results of our work that out of the three distances, the algorithm produces the fastest as well as the most expected result when Euclidean distance has been considered and the slowest as well as the least expected one when Canberra distance has been considered (see analysis in Section 4).

Acknowledgements

The author would like to offer his sincere gratitude to Hemanta Kumar Baruah, Professor, Department of Statistics, Gauhati University, Guwahati, India for his guidance and help in preparation of this article.

References

[1] L. A. Zadeh, “Fuzzy Sets”, Information and Control, vol. 8, Issue 3, **(1965)**, pp. 338-353.
[2] H. J. Zimmermann, “Fuzzy Set Theory and its Applications”, Second Edition, Kluwer Academic Publishers, Boston Massachusetts, **(1991)**.
[3] G. W. Dewit, “Underwriting and Uncertainty”, Insurance: Mathematics and Economics, vol. 1, Issue 4, **(1982)**, pp. 277-285.
[4] J. Lemiare, “Fuzzy Insurance”, Astin Bulletin, vol. 20, Issue 1, **(1990)**, pp. 33-55.

[5] K. Ostaszewski, "An Investigation into Possible Applications of Fuzzy Sets Methods in Actuarial Science", Society of Actuaries, Schaumburg, Illinois, **(1993)**.
[6] L. Zheng and X. He, "Classification Techniques in Pattern Recognition", Conference Proceedings of 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, ISBN 80-903100-8-7 WSCG, Science Press ,Australia, **(2005)**, pp. 77-88.
[7] R. A. Derrig and K. M. Ostaszewski, "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification", Journal of Risk and Insurance, vol. 62, Issue 3, **(1995)**, pp. 447-482.
[8] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, **(1981)**.
[9] J. C. Bezdek and S. K. Pal, "Fuzzy Models for Pattern Recognition: Methods that Search for Structure in Data", IEEE Press, New York, **(1992)**.
[10] S. S. Chauhan, "Advanced Educational Psychology", Fifth Edition, Vikas Publishing Home Pvt. Ltd, New Delhi, **(1993)**.