Emblematic Fuzzy C-means Clustering for Demographic Dataset

Ruchi Arya
TFO-IT
Bangalore, India
aryaruchi25@gmail.com

Abstract—Clustering algorithms are very useful in the field of marketing, education, and healthcare and engineering. It has several advantages such as -it deals with different type of data (text, numerical, image, categorical.),handle outliers, fuzzy data and noise, discover the irregular shape of clusters, produces the result that are easily understandable and insensitive to order of input data. This research work proposed an Emblematic Fuzzy C Means algorithm and analyzed its performance on demographic data set which contains 64 countries and 23 attributes and is taken from World Health Organization. For grouping the countries two attributes Human development index (HDI) and control of corruption is taken into consideration. HDI is the comparative measure of health, education and life of expectancy. Control of corruption defines the economical growth. A comparative analysis of conventional Fuzzy C Means (FCM) and Emblematic Fuzzy C Means (EFCM) has been performed to find out best cluster size with maximum performance validity index.

Keywords-Fuzzy C Means Clustering; Emblematic fuzzy C Means clustering; Human Development Index; Demographic Dataset.

I. INTRODUCTION

In the present scenario technology is progressing very fast. The amount of data in database is quickly increasing and new applications are represented by high dimensional feature vectors .So new application requires retrieval and storage of large objects. Finding the useful data in those databases is a very difficult task. Data Mining techniques are useful for analyzing large amount of data [1-2]. Among them, Clustering is one of the important techniques used for analyzing large amount of data. Clustering is the process of dividing the dataset into small groups so that data into same group have the high similarity in comparison to other groups or we can say that clustering provides the simple representation of dataset by dividing it into homogeneous and distinct sub-groups known as clusters. Clustering is dissimilar to classification since it has no predefined classes [1-8].

Today, very large unlabeled datasets are available. Large dataset does not store in memory of typical computer. Fuzzy clustering algorithms are known to be very useful on small to medium-size data sets. Fuzzy C Means is an efficient clustering approach which allows overlapping of clusters with different degree of membership. For improving the performance of fuzzy c means many extensions has been proposed [9-21]. In this present research work an improvement over fuzzy c means clustering has been proposed by using two different ways. Firstly, the performance of clusters can be validated by using performance indices and best cluster number can be evaluated with maximum accuracy. Secondly an improvement is proposed for the dissimilarity measure to have better performance. The extended FCM algorithm is named as Emblematic Fuzzy C Means (EFCM). EFCM clustering is applied on Demographic dataset and countries have three categories like under developed, developing and developed ones. Some countries may belong to two groups. Two attributes Human Development Index and control of corruption have been taken for clustering. Human development index (HDI) is the comparative measure of standards of living, health, life expectancy, and education for countries. This is an important attribute for partitioning the countries whether the country is under developed, developing or developed. Economic policies effect on the standards of living is measured by HDI. The barrier in the development of countries is corruption. Countries having lower corruption rate have better economical condition. Causes for the corruption are poverty, low or poor salary, leadership problems, low risk detection and punishment, low level of technology advancement, high level of insecurity etc.

The paper is divided into seven sections: section 2 and section 3 discuss the related work and Fuzzy C-Means Clustering. The formulation of Emblematic Fuzzy C-means Clustering is proposed in section 4. Section 5 provides the brief description of the dataset and proposed methodology. Experimental results are analyzed in section 6 and finally conclusion is drawn in section 7.

II. RELATED WORK

Ohn utilized the concept of "cluster centers" on categorical dataset [23]. The validity of the proposed clustering algorithm was tested on two well-known datasets-soybean disease and nursery databases. In another research work, Prodip proposed a simple single pass FCM algorithm that neither uses any complicated data structure nor any complicated data compression techniques, yet produces data partitions comparable to fuzzy c means [15]. He performed experiment on five benchmark datasets and found that single pass FCM performed better as compared to FCM in terms of speed. Wang suggested a Global FCM (GFCM) clustering algorithm which was an incremental approach to hi clustering [18]. The aim of GFCM was to solve the problem of being sensitive to initial conditions and local minimum results of FCM. In order to speed up the converging process, he proposed a fast GFCM which produced more satisfactory results by escaping from the sensibility to initial value and improving the accuracy of clustering. Another research work proposed Hybrid Genetic Algorithm by combining Genetic Algorithm with simulated annealing algorithm and applied it to FCM in order to overcome the locality and the sensitivity to initial clustering central [13]. Jian introduced a new clustering algorithm termed as general c-means for the selection of appropriate parameter for clustering algorithm [21]. In this paper, the mean is extended to general mean and a connection between c-mean clustering and fuzzy c-mean is established. Yadav et al. proposed a foggy k-means clustering for lung cancer patients [3]. They have performed the experiments on real data and analyzed that their proposed algorithm performed better as compared to k-means clustering. In another research paper a fuzzy U nearest algorithm is introduced [21]. In this algorithm U nearest neighbor concept is used to restrict membership for removing noises, isolated point and uninterested data. This algorithm first initialize the cluster number and cluster centers and then update membership function and cluster centers and delete very small clusters. It is determined by the given sample percent and finally if the new cluster satisfies the demand of membership and sample percent then add the new clusters. It is more efficient than k means and fuzzy c means especially with raw dataset which included lots of noises, uninterested data and isolated data points. In another approach an improved Squzeer algorithm for categorical data clustering is introduced [27]. The previous clustering algorithm focus on the numerical data only however much of data is exit in the categorical form. In this an improved clustering algorithm is introduced and tried to improve clustering accuracies of existing categorical data clustering.

III. FUZZY C-MEANS CLUSTERING

Fuzzy Clustering is a powerful approach used in machine learning. There is variety of Fuzzy Clustering methods available which utilizes different types of distances for cluster membership calculation [28]. Fuzzy C Means (FCM) was pioneered by J. C. Bezdek known as one of the popular approaches of Fuzzy Clustering which uses fuzzy weights based on reciprocal distance. Fuzzy weights decrease the total weighted mean-square error. In this algorithm each data point fits in to clusters with some degree of belongingness. Distance calculation is important in Fuzzy C Means. Distance between data point and cluster centre is calculated as its inverse has been taken as taken as degree of belongingness. Total sum of degree of belongingness for any data point should be one. The membership function in FCM could be defined as follows:

$$\sum_{i=1}^{c} \mu_{ij}=1 \text{ (j=1,2,….n)}, \ \mu_{ij}\in[0,1] \text{ (i=1,2,….c j=1,2,….,n)} \quad (1)$$

Here 'μ_{ij}' represents the membership of i^{th} data point within j^{th} cluster center. 'n' denotes the set of data points. Suppose v represents the set of centers and 'v_j' is center of j^{th} cluster. Initially 'c' cluster center has been selected randomly. Fuzzy membership 'μ_{ij}' is calculated as follows:

$$\mu_{ij}=\mathbf{1}/\sum_{k=1}^{c}(d_{ij}/d_{ik})^{(2/m-1)} \quad (2)$$

Fuzzy centers *'v_j'* can also be computed by using:

$$v_j=(\sum_{i=1}^{n}(\mu_{ij})^m x_i/\sum_{i=1}^{n}(\mu_{ij})^m \ ,\forall j=1,2,…..c \quad (3)$$

Here 'm' represents degree of belongingness and m ∈[1, ∞]. The length between i^{th} data and j^{th} cluster center is represented by d^{ij}. Calculation of v_j to be repeated until the 'objective function j' is minimum [29]. Objective function of FCM is as follows:

$$J(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^m||\boldsymbol{x}_i-v_j||^2 \quad (4)$$

A. Fuzziness Factor

The fuzziness factor 'm' was introduced by Bezdek in 1974 [28].It is also called 'fuzzifier'. When the value of is closed to 1 then it tends to be hard and when the value of m close to the infinity then the clusters tend to the fuzziest state. There is not any theoretical justification for the value of 'm'. By experiment it is seen that when the value of m lies between 1.5 to 3, it gives the better result.

B. Number of cluster 'c'

Number of clusters can vary 2 to infinity according to the data partition desired. Membership functions in fuzzy cluster analysis: As we have discussed earlier that data belongs to each cluster by some degree of belongingness. This represents the fuzzy behavior of the algorithm. A matrix U is build whose members are

between 0 and 1. A smoother line is followed by membership function in FCM. It indicates that an element belong to different clusters with different values of degree of belongingness.

IV. EMBLEMATIC FUZZY C-MEANS CLUSTERING

The important parameter of Fuzzy C Means algorithm is cluster center, degree of fuzziness and distance function. As an extension of classical Fuzzy C Means approach following modification has been introduced in Emblematic Fuzzy C Means algorithm:

- In this algorithm a new factor emblematic function (w) and emblematic factor (γ) is introduced.

Emblematic function $w_{ij}=(\|z_j-b_i\|^{-2/(\gamma-1)})/(\sum_{l=1}^{c}\|z_j-b_l\|)^{-2/(\gamma-1)}$ (5)

$\mu_{ij}=(\|z_j-b_i\|^{-m/(m-1)})/(\sum_{l=1}^{c}\|z_j-b_l\|)^{-m/(m-1)}$ (6)

- The cluster center is updated as:

$$b_i=(\sum_{j=1}^{n}.\ \mu_{ij}^{m}.z_j+\sum_{j=1}^{n} w_{ij}^{\gamma}\ z_j)/\ \sum_{j=1}^{n}.(\mu_{ij}^{m}+w_{ij}^{\gamma}) \quad (7)$$

- Now the objective function become:

$$J_{m\gamma}=\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij}^{m}+w_{ij}^{\gamma})\ \|z_j-k_i\|^2 \quad \text{where } 1<m<\infty \text{ and } 1<\gamma<\infty \quad (8)$$

Figure 1 shows different steps involved in Emblematic Fuzzy C Means algorithm.

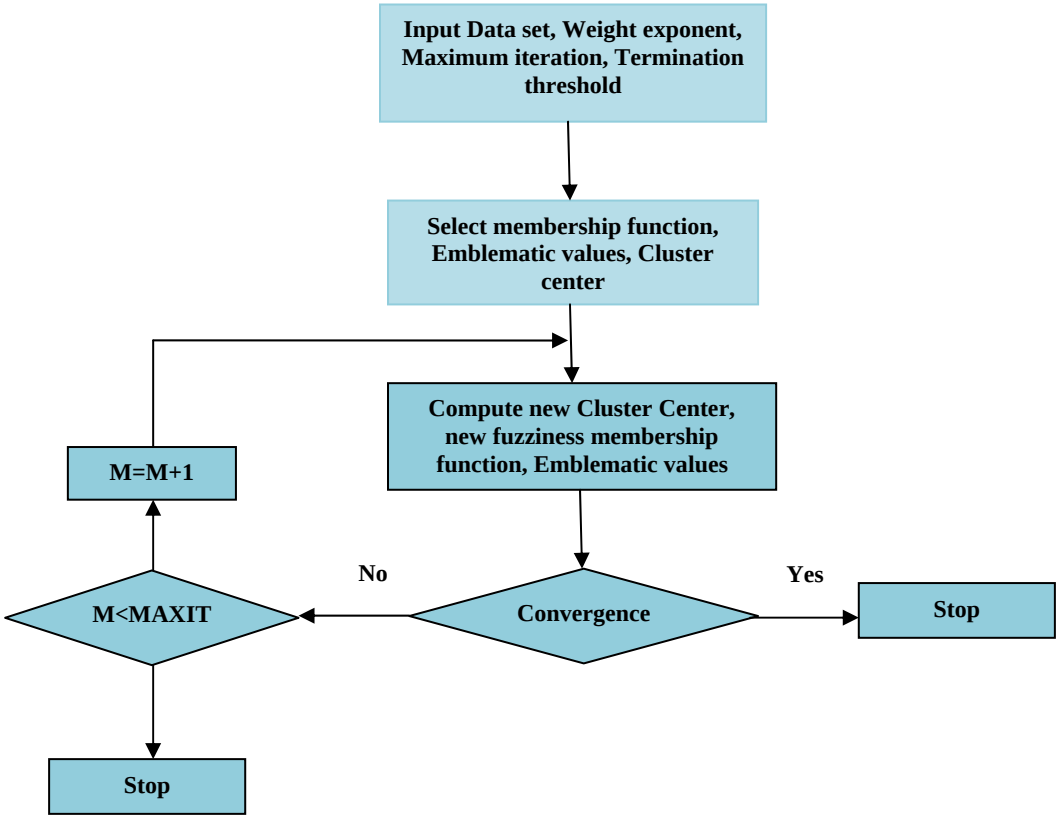

Figure 1. Steps in Emblematic Fuzzy C Means Algorithm

Following are the steps of proposed Emblematic FCM algorithm:

Step 1: Load the dataset in excel format and initialize the input parameters such as-Initial Membership Matrix, Error Constraints, Termination Measures, Weighting Exponent, Number of clusters, Number of Iterations, Emblematic Factor

Step 2: Perform Emblematic Fuzzy C Means Clustering Update membership matrix U and calculate cluster center bi , modified distance dij, Membership function μij, Emblematic Function wij.

Step 3: Test the Model by using the Performance Index Parameters and on the basis of best set of parameters choose the optimum number of clusters.

Step 4: Take the optimum number of cluster and plot clusters.

V. DATASET DESCRIPTION AND PROPOSED METHDOLOGY

A. Dataset Description

Clustering of Demographic data has been performed at global level leading to development of interesting group and for business prospects at different locations across world. A data bank from official website of World Health Organization (WHO) has been referred for Data collection purpose and a demographic data of 64 countries has been taken with 23 attributes for the period of 2001-2006. The Key attributes are Country name, political stability, Human Development Index, Government Effectiveness, GDP per Capita, Life expectancies, GINI, Inflation (GDP Deflator), Population Density (per Sq Km) etc. It is tried to take different country profiles in terms of GDP, life expectancies, GINI, Human development index, Net foreign direct investment inflows, corruption control, population growth to identify the association and correlation using FCM clustering. The Data is collected in Excel format from different tables of the WHO Data Bank. Table 1 shows the brief description of dataset. Human development index (HDI) is the comparative measure of standards of living, health, life expectancy, and education for countries [30-33]. This is an important attribute for partitioning the countries whether the country is under developed, developing or developed. Economic policies effect on the standards of living is measured by HDI. The barrier in the development of countries is corruption. Countries having lower corruption rate have better economical condition. Causes for the corruption are poverty, low or poor salary, leadership problems, low risk detection and punishment, low level of technology advancement, high level of insecurity etc.

TABLE I. DATASET DESCRIPTION

S. No.	Attribute names	Attribute Description
1	Country	Name of the country
2	Press Freedom Index	It is an annual ranking of countries compiled and published by Reporters Without Borders based upon the organization's assessment of their press freedom records
3	Economic Freedom Index	Economic freedom is the fundamental right of every human to control his or her own labor and property.
4	Human Development Index	HDI is a comparative measure of life expectancy, literacy, education and standards of living for countries worldwide.
5	Control of Corruption	Political corruption is the use of legislated powers by government officials for illegitimate private gain.
6	Political Stability	The Political Instability Index shows the level of threat posed to governments by social protest.
7	Government Effectiveness	This index, based on 17 component sources, measures the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.
8	Voice and Accountability	It captures perception of the extent to which a country's citizens are able to participate in selecting their government as well as freedom of expression, freedom of association and free media.
9	Combined Gross Enrollment Ratio for Primary, Secondary and Tertiary schools	It is a statistical measure used in the education sector and by the UN in its Education Index.
10	GDP per Capita	An approximation of the value of goods produced per person in the country, equal to the country's GDP divided by the total number of people in the country.
11	Net Foreign Direct Investment inflows	**It** refers to the net inflows of investment to acquire a lasting management interest in an enterprise operating in an economy other than that of the investor.
12	Female Economic activity rate	It is a measure of women over the age of fifteen who are working

		or able to work as a percent of males.
13	Life Expectancy	Life expectancy is the expected number of years of life remaining at a given age.
14	Religious Freedom	**Freedom of religion** is a principle that supports the freedom of an individual or community
15	Consumer price index	A **consumer price index** (CPI) measures changes in the price level of consumer goods and services purchased by households.
16	GDP growth	It is primarily driven by improvements in productivity, which involves producing more goods and services with the same inputs of labour, capital, energy and materials.
17	Population	Total population of a country.
18	Rural Population	Number of people that are living in rural area.
19	Urban Population	Number of people that are living in urban area.

- **Human Development Index**

HDI is the comparative measure of health and life expectancy for countries. This is an important attribute for categorizing countries in underdeveloped, developing and developed. It measures effect of economic policies on standard of living.HDI is also for states, village, cities by local organization or companies. For calculating HDI following three indices are used:

$$\text{Life Expectancy Index (LEI)} = \frac{LE-20}{83.2-20} \quad (9)$$

$$\text{Education Index (EI)} = \frac{\sqrt{MYSI.EYSI}-0}{0.951-0} \quad (10)$$

$$\text{Mean Year of schooling Index (MYSI)} = \frac{MYS-0}{13.2-0} \quad (11)$$

$$\text{Expected Years of schooling Index (EYSI)} = \frac{EYS-0}{20.6-0} \quad (12)$$

$$\text{Income Index (II)} = \frac{\ln(GNIpc)-\ln(163)}{\ln(108.211)-\ln(163)} \quad (13)$$

WE find HDI by taking the geometric mean of above given indices:

$$HDI = \sqrt[3]{LEI.\,EI.\,II} \quad (14)$$

Countries have higher HDI come under developed countries, countries having medium value come under developing countries and countries having lower value of HDI come under developing countries.

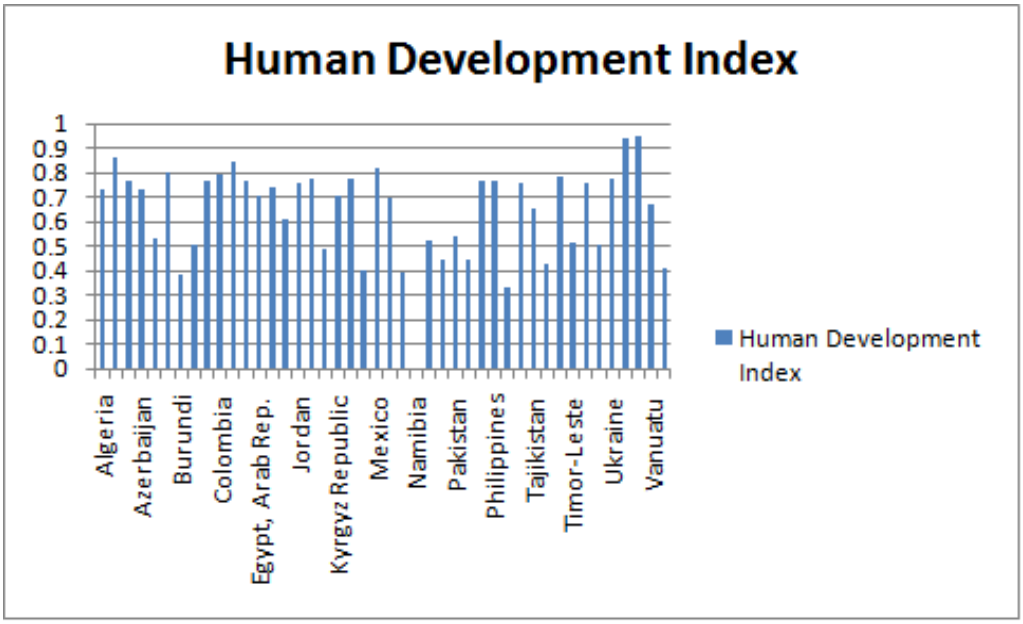

Figure 2. HDI of different countries

- **Control of Corruption**

Corruption is obstacle for development of countries. Countries having lower corruption rate have better economical condition. Causes for the corruption are poverty, low or poor salary, leadership problems, low risk detection and punishment, low level of technology advancement, high level of insecurity etc. Control of corruption means the countries and state prevent public servants and politician from taking bribes, extortion, nepotism, cronyism, graft and embezzlement.

- **Life Expectancy**

Life expectancy is commonly utilized and analyzed component for demographic data for the countries of world. It represents the overall health of a country. It can fall due to war, disease, poor health and food crisis. Life expectancy can be improved by improving health, happiness, nutrition and medicine increase. The countries having higher life expectancy are developed countries. Countries having high GNP does not mean that the country have high life expectancy.

- **Gross Domestic Product**

Standard of living different countries is generally indicated by GDP. GDP can be determined by Product, Income and Expenditure approach. The product approach is the direct approach for determining the GDP. Base of expenditure approach is that all the products brought by somebody. So the total product value will be equal to total expenditure in buying things. The GDP is determined by adding all the producers' income. GDP per capita is equal for comparing the countries because it shows relative performance of countries. Increase in this shows growth in economy and increase in productivity. Countries having the higher degree of living are come under the developed countries.

- **Economic Freedom**

It is right of every human. It shows the human control of his property and effort. In economically free society everybody is free to produce, work, consume and invest in which manner they want. For determining the economic freedom there are ten components:

i. Business Freedom
ii. Investment freedom
iii. Trade freedom
iv. Financial freedom
v. Fiscal freedom
vi. Property rights
vii. Government Spending
viii. Freedom from Corruption
ix. Monetary Freedom
x. Labor Freedom

B. Proposed Methodlogy

Figure 3 shows the proposed methodology. First we take the demographic dataset and on this dataset Fuzzy C means and Emblematic fuzzy c means algorithm are applied. Different performance parameters are calculated in order to compare their performance.

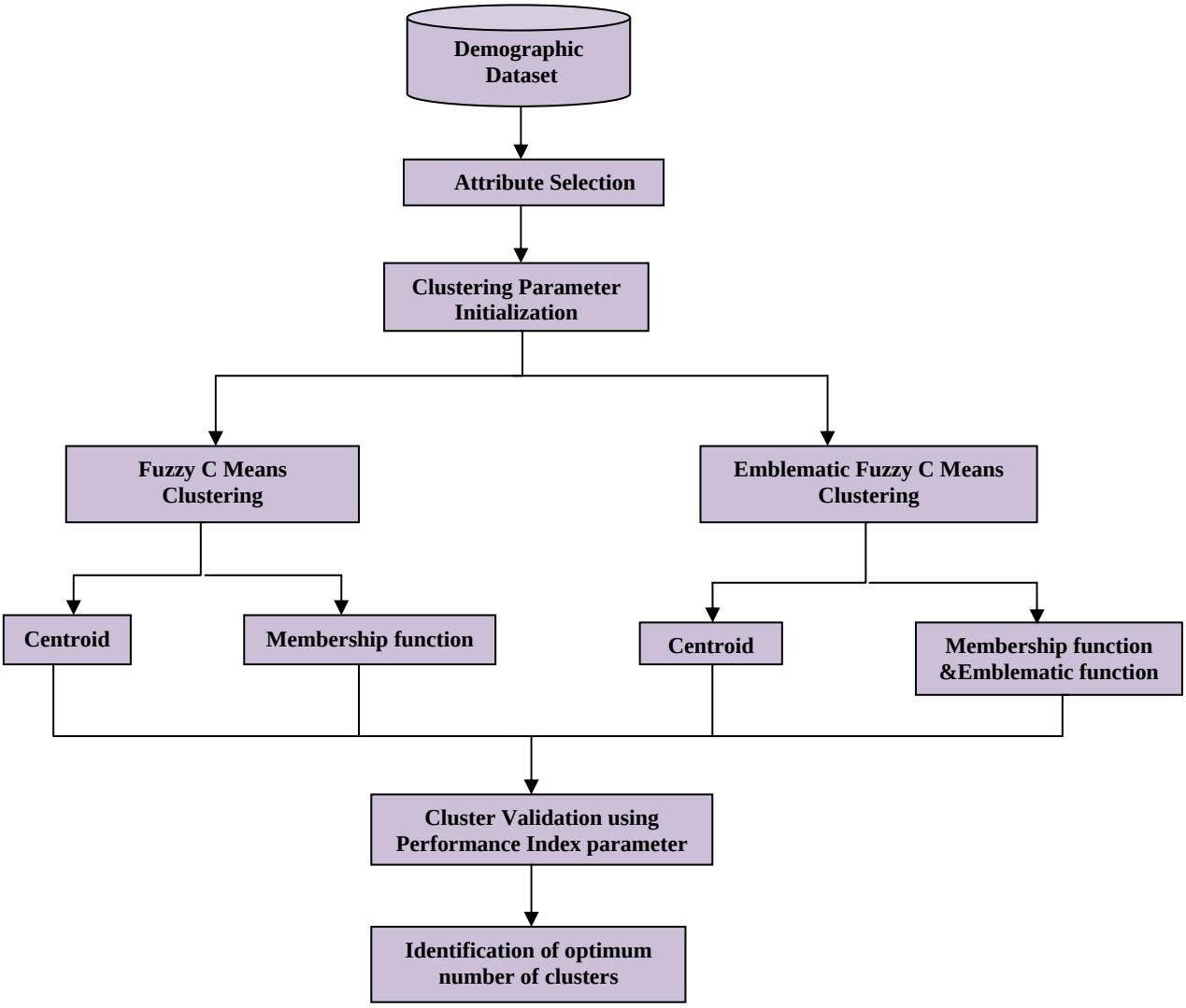

Figure 3. Proposed Methdology

C. Validity Index for Fuzzy Clustering

The Validity functions help us to validate the correct number of clusters for clustering algorithm [28]. Some famous validity indexes are described as follows:

- ***Partition Coefficient:*** First validity index associated with FCM is Partition Coefficient (PC). It measures the overlapping area between clusters defined as:

$$PC(c)=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{2} \quad (15)$$

  Where PC(c) lies between 1/c and 1.For finding the optimal cluster number c* maximum value of PC(c) is taken ,c lies between 2 to (n-1) to produce a best clustering performance for the data set Z.

- ***Partition Entropy:*** It is defined as:

$$PE(c)=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\log_2\mu_{ij} \quad (16)$$

  where PE(c) lies between 0 to log2c. For finding the optimal cluster the minimum value of PE(c) is taken where c lies between c to (n-1). It will give best clustering performance.

- ***Xie and Beni's Index:*** Xie and Beni's Index (XB) is used to to calculate optimal number of clusters [34]. It is illustrated by the fraction of the sum of inner variation within clusters and their separation. Minimum value of XB index indicates optimal number of clusters.

$$XB(c)=(\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}||x_j-v_i||^2/(n\ \min_{ij}||x_j-v_i||^2) \quad (17)$$

- ***Partition Index (SC):*** Partition Index (SC) is one of the important cluster validation parameters and denoted as fraction of sum of dense or compact clusters and their inter cluster separation [31]. In Partition Index sum of dense or compact clusters should be normalized by divided it with the fuzzy cardinality values of each cluster.

$$SC(c)=\sum_{i=1}^{c}(\sum_{j=1}^{N}(\mu_{ij})^{m}||x_j-v_i||^2)/(N_i\sum_{k=1}^{c}||v_k-v_i||^2 \quad (18)$$

  Partition index is useful to identify best partition when each groups having same number of clusters. A lesser value of partition index is desirable to have a better partition.

- ***Separation index(S):*** Separation index in contrast of partition index, uses the minimum distance separation for partition validity.

$$S(c)=(\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{2}||x_j-v_i||^2)/(n\ \min_{i,k}||v_k-v_i||^2) \quad (19)$$

- ***Dunn's Index:*** This metric show the compact separated (CS) clusters.

$$DI(c)=\min_{i\varepsilon c}\{\min_{j\varepsilon c,i\neq j}\{\min_{x\varepsilon c1,y\varepsilon cj}d(x,y)/\max_{k\varepsilon c}\{\max_{x,,y\varepsilon c}d(x,y)\}\} \quad (20)$$

- ***Alternative Dunn Index(ADI):*** The high value of DI indicates dense as well as well separated clusters. The calculation of DI becomes computationally expensive with higher value of cluster centers and data points. An improvement over Dunn Index was proposed as Alternative Dunn Index (ADI) to reduce the computational complexities. A higher value of ADI has be achieved to identify optimum number of clusters.

$$ADI(c)=\min_{i\varepsilon c}\{\min_{j\varepsilon c,i\neq j}\{(\min_{xi\varepsilon Ci,xj\varepsilon cj}\left|\ d(y,v_j)-d(x_i,v_j)\ \right|)/\max_{k\varepsilon c}\{\max_{x,y\varepsilon c}d(x,y)\}\} \quad (21)$$

Optimal Number of Clusters is required to mention before starting of the clustering process and it is difficult to estimate. Few validity parameters defined above for evaluating the cluster performance and they can also provide an idea about number of clusters required. Here, it is required to calculate all above cluster validation parameters and a cluster set with optimal values of validation parameter may be chosen.

VI. EXPERIMENT AND DISCUSSION

Experiment is done by using Matlab 7.0. In this chapter, FCM and EFCM algorithm are performed on the demographic dataset and different performance validity indices such as Partition co-efficient (PC), partition entropy (PE), Xie and Beni's Index (XB), Separation index(S) DI and ADI are calculated. For a valid cluster number PC, DI, ADI should be maximum and CE, S, XB and SC should be minimum. From our experiment cluster number 3 verify all the conditions. So for cluster number 3 EFCM gives the good clustering result.

A. Cluster Center Initialization of FCM Clustering on Demographic Dataset

When fuzzy c means clustering is performed on small dataset it performs well. But when it is performed on Demographic dataset it gives unusual results. The initial cluster centers coincide with each other.

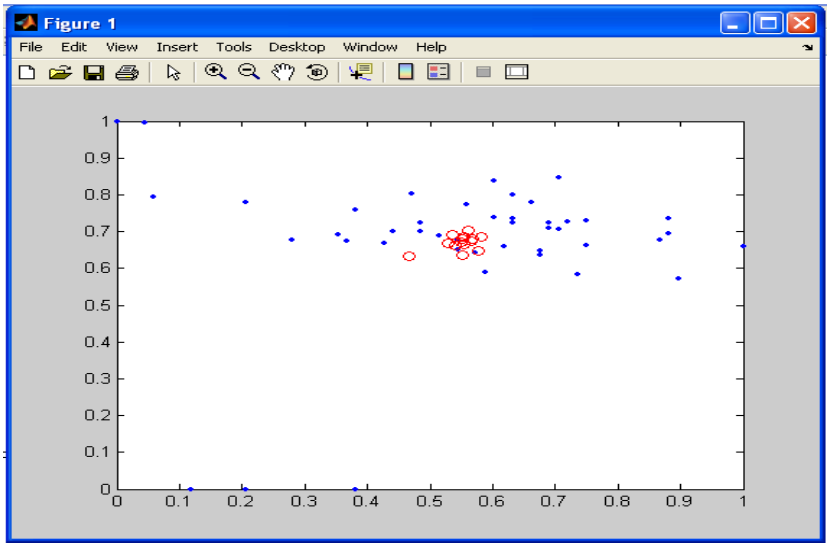

Figure 4. Cluster center initialization of FCM Clustering on Demographic Dataset

B. Different Performance Indices for FCM clustering on Demographic Dataset

The different performance indices when conventional FCM clustering is performed on Demographic dataset are in Table 2.

TABLE II. PERFORMANCE INDICES FOR DIFFERENT NUMBER OF CLUSTERS OF FCM ALGORITHM

# clusters	PC	CE	SC	S	XB	DI	ADI
2	0.9088	0.1574	1.1049e-008	2.5111e-010	8.2059	0.0344	0.0049
3	0.9805	0.0427	2.0408e-010	.0701e-012	25.1428	0.6761	0.0940
4	0.9200	0.1565	7.9184e-011	7.7904e-0122	8.8484	0.1011	0.0270
5	0.8146	0.3531	2.6240e-010	8.3098e-012	7.5190	0.0122	0.0264
6	0.8587	0.2879	6.2004e-011	2.1219e-012	4.8526	0.2975	0.0194
7	0.7592	0.5121	1.4660e-010	5.2405e-012	5.1457	0.0161	0.0134
8	0.8114	0.3914	5.7414e-010	1.5793e-011	7.5900	0.0301	0.0032
9	0.8138	0.4116	2.8156e-010	9.0310e-012	3.8271	0.0253	0.0141
10	0.8164	0.4059	8.1528e-011	2.8202e-011	5.5695	0.0301	0.0032
11	0.6914	0.6466	4.1638e-010	1.2099e-011	10.3312	0.3781	0.0240
12	0.6082	0.8493	1.8956e-010	6.2227e-012	4.0532	0.0126	0.0024
13	0.7736	0.5309	6.9816e-011	2.1643e-012	18.9460	0.0160	0.0026
14	0.8139	0.3839	6.2869e-010	1.7790e-011	10.6565	0.1270	0.0020

C. Cluster Center Intialization of EFCM Clustering on Demographic Dataset

Figure 5 shows the initial cluster center of demographic dataset when emblematic fuzzy c means algorithm is performed.

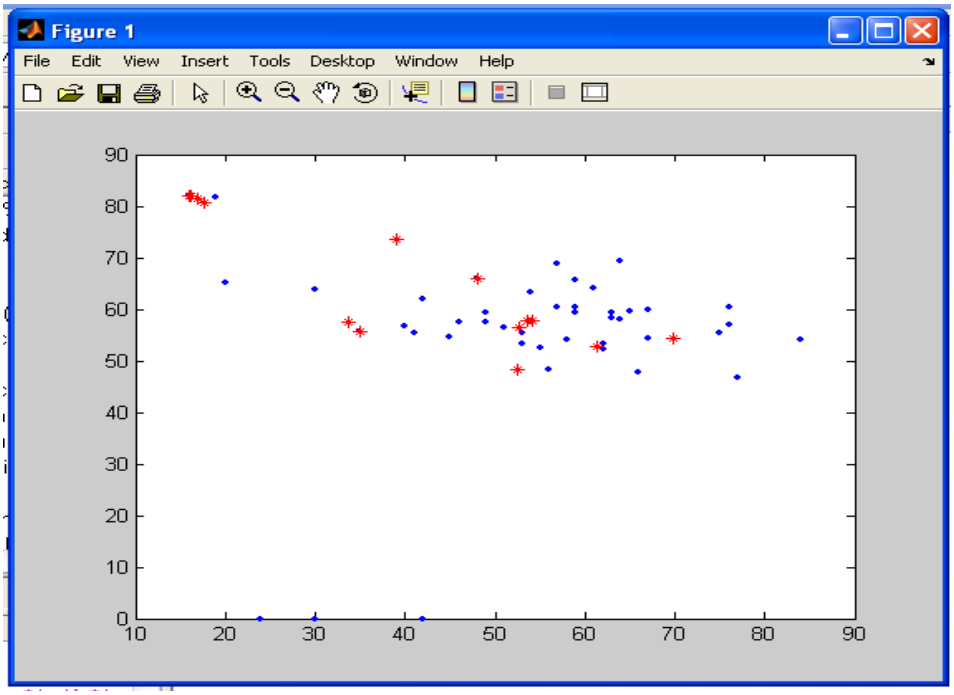

Figure 5. Cluster center Initialization of EFCM Algorithm on Demographic Dataset

D. Different Performance Indices of EFCM clustering on Demographic Dataset

Different performance indices when EFCM clustering algorithm is performed on Country dataset are in Table 3.

TABLE III. PERFORMANCE INDICES FOR DIFFERENT NUMBER OF CLUSTERS OF EFCM ALGORITHM

# clusters	PC	CE	SC	S	XB	DI	ADI
2	0.9492	0.0811	1.3924e-008	3.1645e-010	14.2331	0.0163	0.0048
3	0.9991	0.0028	6.8391e-011	1.9503e-012	4.2399	0.6761	0.0466
4	0.9720	0.0432	5.7922e-011	1.5397e-012	49.9927	0.8832	0.0174
5	0.9248	0.1386	2.9499e-010	8.3207e-012	6.2036	0.0110	0.0243
6	0.9633	0.0661	7.5660e-011	2.1642e-012	28.2363	0.1226	0.0120
7	0.8465	0.3051	9.6974e-011	3.7621e-012	19.8845	0.0032	0.0027
8	0.9438	0.1042	5.6773e-010	1.3868e-011	7.4988	0.1273	0.0099
9	0.9609	0.0806	1.8882e-010	5.7513e-012	16.4086	0.2975	0.0138
10	0.9482	0.0948	3.2310e-010	9.9638e-012	42.1812	0.1549	0.0110
11	0.9416	0.1103	2.2005e-010	5.5808e-012	10.2861	0.0312	0.0170
12	0.9336	0.1499	2.0250e-010	1.0707e-011	4.56150	0.0152	0.0593
13	0.9104	0.2010	5.7792e-011	1.5861e-012	10.0848	0.0353	0.0075
14	0.9366	0.1081	6.6043e-010	1.7911e-011	6.0851	0.0469	0.0162

Countries are grouped into three clusters. Some countries may belong to more than a cluster. In our experiment countries can be categorize into developed, developing and underdeveloped. Two attributes have been taken into consideration: Human development index and corruption control .Countries having high human development index and corruption control came in developed countries. Some countries are at the border they can belong to the both the groups. Table 4 defines the three different clusters of countries.

TABLE IV. DENTIFIED CLUSTERS AND THEIR COUNTRIES

Cluster Names	**Countries**
Developed Countries	United States, United Kingdom, Czech Republic, Australia, Canada, Denmark, Germany, France, Hongkong , Japan, Italy, Newzealand, Singapore
Developing countries	Newzealand , Denmark, Czech Republic, Italy , India, Colombia, Libya, Malawi, Kenya, Mozambique, Moldova , Sri Lanka , Nepal , Philippines ,Peru ,Zimbabwe ,Pakistan, Algeria , Argentina , Bangladesh , Bhutan , Afghanistan , Brazil, Bosnia and Herzegovina, China , Cameroon , Malaysia, Maldives, Mexico ,Thailand ,Tanzania, Uganda, Turkey, Ecuador ,Egypt , Ghana, Georgia, Namibia.
Under developed countries	Bangladesh, Mozambique, Afghanistan, Chad, Burundi, Angola, Guinea, Myanmar , Niger , Gambia, Liberia , Malawi, Zambia, Timor-Lessee, Bhutan, Tajikistan, Sierra Lone, Yemen, Sudan, Togo, Senegal.

In the above experiment Newzealand, Italy, Denmark and Czech Republic came in both developed and developing countries. Bangladesh, Bhutan, Afghanistan Malawi and Mozambique belong to both developing and underdeveloped countries. In figure 6 shows that the PC is maximum for cluster number three and CE is minimum for cluster number three which is essential.

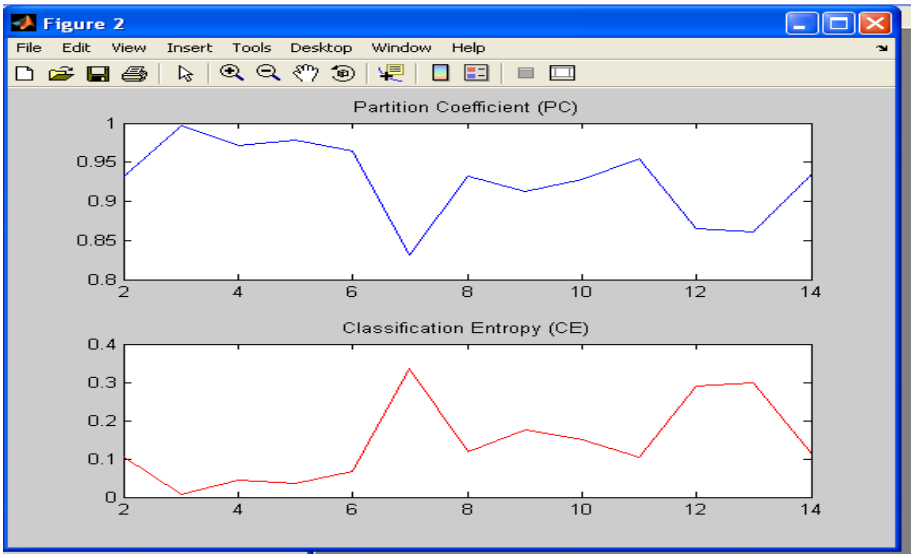

Figure 6. Partition Coefficient (PC)and Classification Entropy (CE)of EFCM on Demographic Dataset

In figure 7 indicates that the partition index, separation index are almost constant after cluster number three. XB is almost constant for cluster number 13 and after that it increases. From these we see that for cluster number three PC is maximum and CE,S, XB are minimum which are required. In the same way, figure 8 shows the DI, which is maximum for cluster number 3.

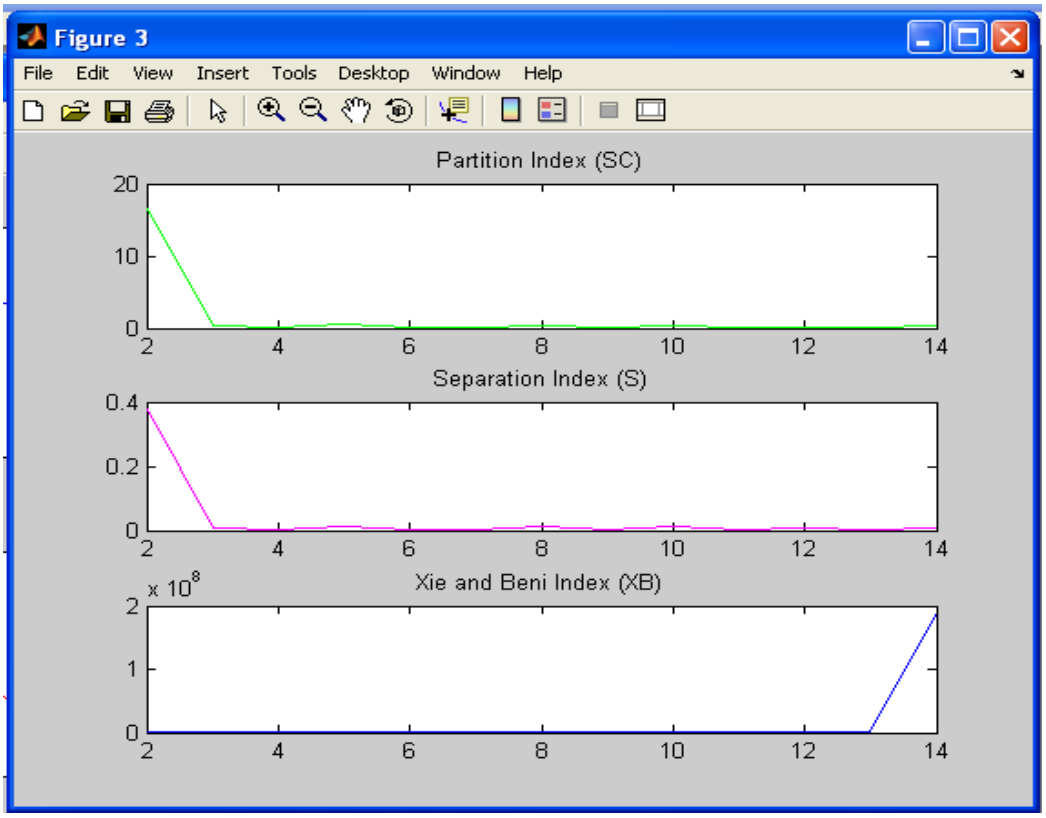

Figure 7. Partition Index(SC),Separation Index(S)and Xie and Beni Index(XB)of EFCM on Demographic Dataset

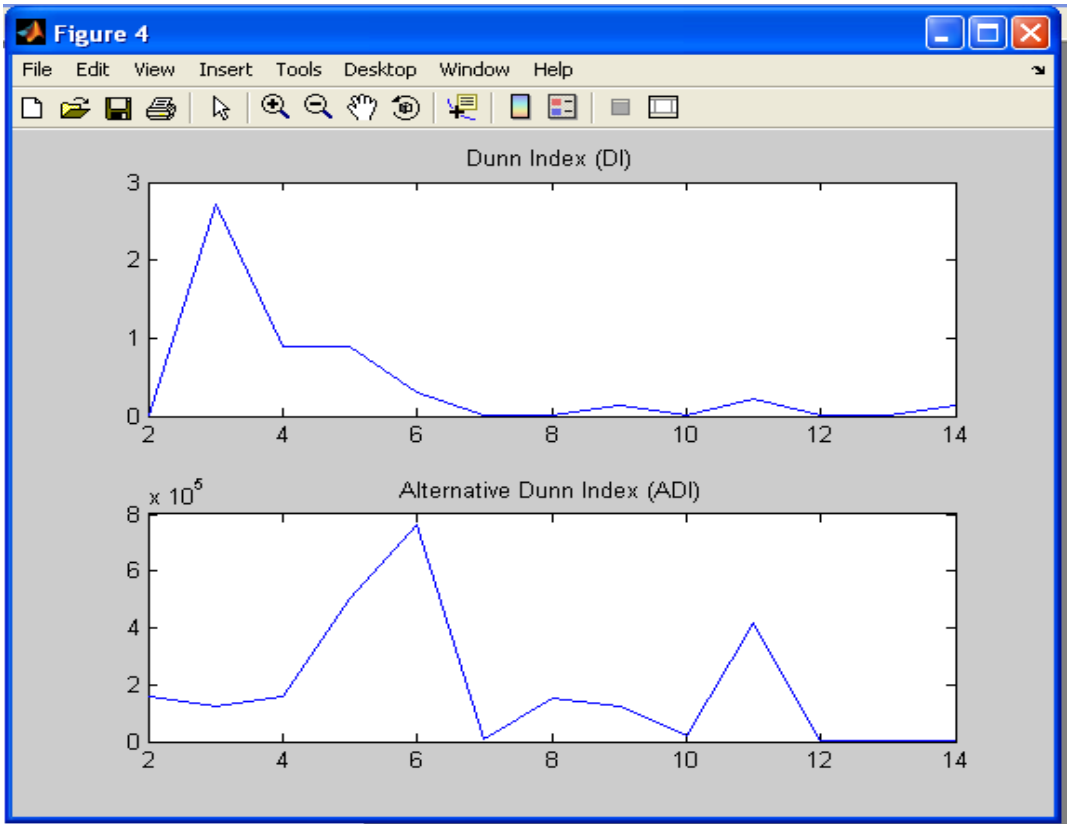

Figure 8. Dunn Index(DI) and Alternative Dunn Index(ADI) of EFCM on Demographic Dataset

VII. CONCLUSION

In Fuzzy C-Means Clustering, when the dataset is large the initial cluster centers coincide with each other. For removing this shortcoming of Fuzzy C Means algorithm, this research work has added some extra term, also called emblematic factor, in classical FCM algorithm. The performance of different clustering algorithms is evaluated using Partition co-efficient (PC), Partition Entropy (PE), Xie and Beni's Index (XB), Separation Index(S), Dunn Index (DI) and Alternative Dunn Index (ADI). For a valid cluster number PC, DI, ADI should be maximum and PE, S, XB and SC should be minimum. The performance of conventional FCM and proposed EFCM is evaluated on the demographic dataset. From the experimental work, it is analyzed that for three clusters all the performance index values are verified and the proposed EFCM has obtained better performance indices as compared to traditional FCM.

REFERENCES

[1] Han, Jiawei, and Micheline Kamber. Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann, 2006.
[2] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science & Bio-Technology 5, no. 5 (2013).
[3] Yadav, Akhilesh Kumar, Divya Tomar, and Sonali Agarwal. "Clustering of lung cancer data using Foggy K-means." In Recent Trends in Information Technology (ICRTIT), 2013 International Conference on, pp. 13-18. IEEE, 2013.
[4] Bensaid, Amine M., Lawrence O. Hall, James C. Bezdek, Laurence P. Clarke, Martin L. Silbiger, John A. Arrington, and Reed F. Murtagh. "Validity-guided (re) clustering with applications to image segmentation." Fuzzy Systems, IEEE Transactions on 4, no. 2 (1996): 112-123.
[5] Bezdek, James C. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, 1981.
[6] Agarwal, Sonali. "Classification of Countries based on Macro-Economic Variables using Fuzzy Support Vector Machine." International Journal of Computer Applications 27, no. 6 (2011).
[7] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999): 264-323.
[8] Rathore, Neha, and Sonali Agarwal. "Predicting the survivability of breast cancer patients using ensemble approach." In Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, pp. 459-464. IEEE, 2014.
[9] Agarwal, Sonali, and Divya Tomar. "A Feature Selection Based Model for Software Defect Prediction." International Journal of Advanced Science and Technology,Vol.65, pp.39-58, 2014.
[10] Di Nuovo, S., and Vincenzo Catania. "An evolutionary fuzzy c-means approach for clustering of bio-informatics databases." In Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference on, pp. 2077-2082. IEEE, 2008.
[11] Vanisri, D., and C. Loganathan. "An Efficient Fuzzy Clustering Algorithm Based on Modified K-Means." International Journal of Engineering Science and Technology 2, no. 10 (2010): 5949-5958.
[12] Jiang, Lei, and Wenhui Yang. "A modified fuzzy c-means algorithm for segmentation of magnetic resonance images." In Proc. VIIth digital image computing: Techniques and applications. 2003.

[13] Su-hua, Liu, and Hou Hui-fang. "A combination of mixture Genetic Algorithm and Fuzzy C-means Clustering Algorithm." In IT in Medicine & Education, 2009. ITIME'09. IEEE International Symposium on, vol. 1, pp. 254-258. IEEE, 2009.
[14] Cannon, Robert L., Jitendra V. Dave, and James C. Bezdek. "Efficient implementation of the fuzzy c-means clustering algorithms." Pattern Analysis and Machine Intelligence, IEEE Transactions on 2 (1986): 248-255.
[15] Hore, Prodip, Lawrence O. Hall, and Dmitry B. Goldgof. "Single pass fuzzy c means." In Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International, pp. 1-7. IEEE, 2007.
[16] Agarwal, S. "Weighted support vector regression approach for remote healthcare monitoring." In Recent Trends in Information Technology (ICRTIT), 2011 International Conference on, pp. 969-974. IEEE, 2011.
[17] Tomar, Divya, Shubham Singhal, and Sonali Agarwal. "Weighted Least Square Twin Support Vector Machine for Imbalanced Dataset." International Journal of Database Theory & Application 7, no. 2 (2014).
[18] Wang, Weina, Yunjie Zhang, Yi Li, and Xiaona Zhang. "The global fuzzy c-means clustering algorithm." In Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on, vol. 1, pp. 3604-3607. IEEE, 2006.
[19] Tomar, Divya, Ruchi Arya, and Sonali Agarwal. "Prediction of profitability of industries using weighted SVR." International Journal on Computer Science and Engineering 3, no. 5 (2011): 1938-1945.
[20] Wang, Xiao Ying, Jon Garibaldi, and Turhan Ozen. "Application of the fuzzy C-means clustering method on the analysis of non pre-processed FTIR data for cancer diagnosis." In Internat. Conf. on Australian and New Zealand Intelligent Information Systems (ANZIIS), pp. 233-238. 2003.
[21] Wang, Yiding, and Qiaona Pei. "Fuzzy U Nearest Neighbor Adaptive Clustering Algorithm." In Computer Science and Software Engineering, 2008 International Conference on, vol. 6, pp. 189-192. IEEE, 2008.
[22] Nath, Sayantan, Sonali Agarwal, and Qasima Abbas Kazmi. "Image histogram segmentation by multi-level thresholding using Hill climbing algorithm." International Journal of Computer Applications 35, no. 1 (2011).
[23] San, Ohn Mar, Van-Nam Huynh, and Yoshiteru Nakamori. "An alternative extension of the k-means algorithm for clustering categorical data." International Journal of Applied Mathematics and Computer Science 14 (2004): 241-247.
[24] Yu, Jian. "General c-means clustering model and its application." In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 2, pp. II-122. IEEE, 2003.
[25] Tomar, Divya, and Sonali Agarwal. "Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease." International Journal of Bio-Science & Bio-Technology 6, no. 2 (2014).
[26] Agarwal, Sonali, and G. N. Pandey. "SVM based context awareness using body area sensor network for pervasive healthcare monitoring." In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, pp. 271-278. ACM, 2010.
[27] He, Zengyou, Xiaofei Xu, and Shenchun Deng. "Improving categorical data clustering algorithm by weighting uncommon attribute value matches." Computer Science and Information Systems 3, no. 1 (2006): 23-32.
[28] Bezdek, James C. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, 1981.
[29] Bensaid, Amine M., Lawrence O. Hall, James C. Bezdek, Laurence P. Clarke, Martin L. Silbiger, John A. Arrington, and Reed F. Murtagh. "Validity-guided (re) clustering with applications to image segmentation." Fuzzy Systems, IEEE Transactions on 4, no. 2 (1996): 112-123.
[30] http://www.sginetwork.org/index.php?page=indicator_quali&indicator=S4_4
[31] http://lukemcbain.wordpress.com/2010/07/09/thoughts-on-anti-corruption-and-leadership/
[32] http://geography.about.com/library/weekly/aa042000b.htm
[33] http://www.heritage.org/index/
[34] Xie, Xuanli Lisa, and Gerardo Beni. "A validity measure for fuzzy clustering." IEEE Transactions on pattern analysis and machine intelligence 13, no. 8 (1991): 841-847.