International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 8 (2015) pp. 20629-20644

http://www.ripublication.com

Clubbing of Fuzzy Relational Data Sets into Heterogeneous Cluster Bags using Heuristic Multi Domain Fuzzy Clustering Technique

P.Sivasankari[1], K.S.Ravichandran[2], R.KrishanKumar[3]

[1]Department of Advanced Computing, SASTRA University,Thanjavur-613401, Tamil Nadu, India
[2]Department of Information and Communication Technology, School of Computing, SASTRA University,Thanjavur-613401, Tamil Nadu, India
[3]Research Scholar, School of Computing SASTRA University, Thanjavur-613401, Tamil Nadu, India
Email:sankarimp@gmail.com,raviks.it@sastra.edu, krishankumar@sastra.ac.in

Abstract

In present competitive world data is everything. Man and machines interact with each other just because of this data governing them. People have subtle time to gather data from the vast spread domain. To overcome such time spares scientists have evolved easy data structures to club data together, which most of us know as Data Set. These data sets have an implicit factor of relationship by themselves, which many of us have rarely admired. There is indeed a narrow stream of literature that has tackled this aspect. To further gear the accelerator, this paper proposes a technique called Multi Domain Fuzzy Cluster (MDFC) to club data sets into bags of Densely Related Cluster, Loosely Dense Related Cluster and Strongly Densely Related Clusters. As the data set taken for study are fuzzy in nature, it is working for a dependency matrix and fed as an input along with the classical dependency matrix to understand the club pattern. The motivation of this technique has emerged from the classical and Fuzzy Functional Dependencies.

Keywords: Multi Domain Fuzzy Cluster, Classical functional Dependency, Fuzzy Functional Dependency.

1. Introduction

Data as a magical spell is a man's best friend. In today's fast moving world data is the fuel for fitness. The world is rich with data and meager with guidelines and

techniques to calibrate them into effective information. Many a times, it so happens that people gain data in abundance and are using them in a very subtle way. The major factor for this ineffective data management is that there is less attention drawn to the implicit relationship that the data poses. There is a wide range of technique available in the literature to handle external relationship of the data both based on context and non-context dependencies. The implicit relationship is a relationship that the data set taken for study forms within itself by relating the attributes of the concerned data set. The relationship between attributes is given by using proximity relation. It can be taken as a constraint for clustering the databases. The proximity gives better values for relationships than other techniques. The proximity clustering provides better results than other clustering [1].Most of the time in approaches like Data Mining the end result to the user suffers because of wrong data set selection. Here wrong signifies the relationship factor between various attributes in the data set. To tackle this issue a versatile technique called the Multi Domain Fuzzy Cluster is adopted. Real time data are ambiguous in nature and so applying a simple technique of classical dependency will alone not serve the purpose, so an emerging need for fuzzy dependency is put forth. Relational Database plays important role in developing a software, analyzing the data and in administrating the process. Presence fuzziness in database leads to transformation. The dependencies in a relational model are a bigger issue. To overcome that conformance technique is applied[2].

This is a problem considered within a data framework and so any type of data set be it a real time or a user framed set is considered for heuristic treatment**.** In order to make this technique effective it is validated using both user framed and real time data. The Multi Domain Fuzzy Cluster (MDFC) is an unique clustering method that is applied to a given data set so as to group them into one of the following cluster bags namely, Densely Cluster (DC) that signifies that the set is a valid candidate for further literature methodology to be applied, Loosely Densely Cluster (LDC) means that literature methodologies can be applied but consideration must be given to remove outliers and redundancies, Strongly Densely Cluster (SDC) means that set is a perfect candidate for the methods to be applied. The technique discussed in this paper is used as an inevitable-pre-analyzer to determine whether the data are getting fit for the purpose or not.

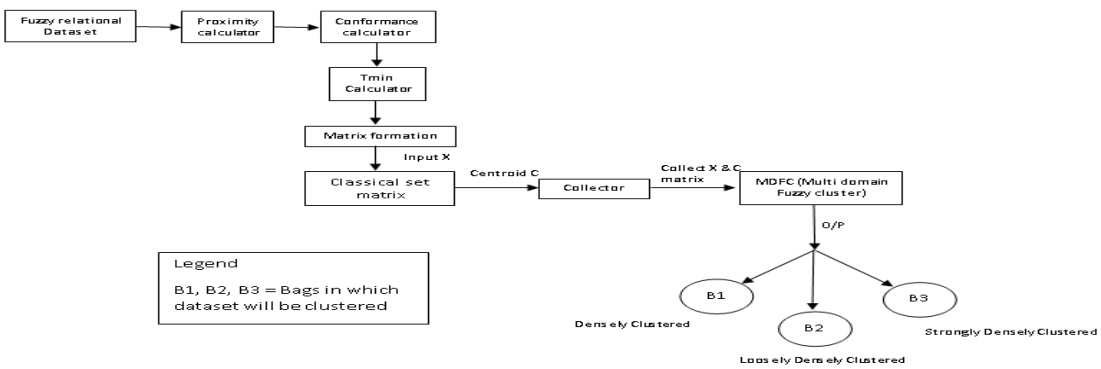

Fig. 1. The proposed Multi Domain Fuzzy Cluster Architecture

The strength of the classical data dependency and Fuzzy data dependency will be utilized in the construction of an efficient clustering technique. Applying fuzzy clustering to linguistic type data set is a tricky thing. This paper proposed the proximity measures for qualitative scales which are not balanced. Here the comparison is made between linguistic variables by adopting a formal approach. This helps in better decision making [3]. The normalizing of linguistic data set is an important aspect of database management. For normalizing of database we use different kinds of Fuzzy Functional Dependencies (FDDs). Here the proposed method is closeness relation. The normalization restricts the anomalies due to insertion and deletions. It splits the relation into two parts. It assures the original relation taken from the partitions. It also employs approximate equality. It produces the degree of closeness that is proximity relation in a relational database model [4].Upon clustering the set into their respective bags the user shall gain confidence of what to do next with the data so that the effort does not end in vain. The understanding of associations along with various linguistic instances is tough. Many techniques have evolved in the recent times. Fuzzy association analysis is the most important part of data mining. Here associations can be expressed in natural language. Linguistic expressions are used to evaluate the association among attributes [5].

The Fig.1 depicts the architecture of the MDFC model. In this model the input is a data set that is fed in after calculating the proximity, conformance, Tmin. The matrix format input is fed along with centroid which is a matrix of classical form. These two matrices are collected as input and fed to MDFC to bag the data set into respective clusters.

1.1 Dataset Description-Heart Disease

The table-1 below shows the Fuzzy Dataset for Cardiac disease that consists of eight attributes which are fuzzy in nature and thereby inherit the fuzziness. The table-2 below consists of the Linguistic Conversion of Fuzzy Cardiac Dataset. The table-3 below shows the Tuples Description. The table-1 has been formulated by collecting real time data set from UCI repository for patients with cardiac disorders in the Cleveland Clinic Foundation of United States

.

Table-1: Fuzzy Dataset for Cardiac Disease

Pid	Age	Sex	CP	BP	Cholesterol	FBS	ECG	Thalach
1	63	1	1	145	233	121	2	150
2	67	1	2	160	286	120	2	108
3	67	1	2	120	229	120	2	129
4	37	1	3	130	250	120	1	187
5	41	2	2	130	204	120	2	172
6	56	1	2	120	236	120	1	178
7	62	2	4	140	268	120	2	160
8	57	2	4	120	354	120	1	163
9	63	1	4	130	254	120	2	147
10	53	1	4	140	203	121	2	155

Table-2: Linguistic Conversion of Fuzzy Cardiac Dataset

Pid	Age	Sex	CP	BP	Cholesterol	FBS	ECG	Thalach
1	Old	Male	Typical Angina	High	Normal	Yes	LVH	Normal
2	Old	Male	Asymptomatic	High	High	No	LVH	Normal
3	Old	Male	Asymptomatic	Normal	Normal	No	LVH	Normal
4	Middle	Male	NonAnginalpain	Normal	High	No	Normal	High
5	Middle	Female	Atypical Angina	Normal	Normal	No	LVH	High
6	Old	Male	Atypical Angina	Normal	Normal	No	Normal	High
7	Old	Female	Asymptomatic	High	High	No	LVH	High
8	Old	Female	Asymptomatic	Normal	High	No	Normal	High
9	Old	Male	Asymptomatic	Normal	High	No	LVH	Normal
10	Middle	Male	Asymptomatic	High	Normal	Yes	LVH	Normal

Table-3: Tuples Description.

Attributes	Description	Conditon range limits
Age	**Age in years** <= 55 – middle >55 -old	55.0
Sex	**Sex** 1 -Male 2 -Female	1.5
CP	**Chest Pain Type** 1 -Typical angina 2 -Atypical angina 3 -Non Anginal pain 4 -Asymptomatic	2.5
BP	**Resting Blood Pressure(in mm Hg)** >=135 -High <135 -Normal	135.0
Cholesterol	**serum cholesterol in mg/dl** <238 -Normal >=238 - High	238.0
FBS	**Fasting blood sugar** > 120 mg/dl 1 = Yes 0 = No	120.0
ECG	**Resting electrocardiographic results** 1-Normal 2-Left Ventricular Hypertrophy(LVH)	1.5
Thalach	**Maximum heart rate achieved** <=155 -Normal >155 -High	155

The further discussion includes Section 2 Methodology, Section 3 Multi Domain Fuzzy Cluster, Section 4 Results and Discussion and Section 5 Conclusion.

2. Methodology

2.1 Classical Data Dependency

The functional dependency in classical relational database determines the existence of functions among two attributes i.e. P and Q. The r is a relation instance on $R(V_1,V_2...,V_n)$, then the U is the universal set of attributes $V_1,V_2...,V_n$. X,Y are the subset of U. We can say that r (relation) holds the functional dependency P→Q(P determines Q) if every pair of tuples t_a, t_b should be $t_a[P] = t_b[P]$ implies $t_a[Q] = t_b[Q]$.

2.2 Fuzzy Data Dependency

The FD (functional dependency) for a fuzzy relational model is defined not by an equality relation but by using a similarity metric. This is because in fuzzy model we have a lot of imprecise values. The extended version of functional dependency is known as Fuzzy Functional Dependency (FFD).

The two fuzzy attributes P and Q are defined under a dependency P→Q(fuzzy (P) determines fuzzy(Q)) if every pair of attributes t_a, t_b should be $t_a[P] = t_b[P]$ implies $t_a[Q] = t_b[Q]$ where (= is not equality but is describing similarity).

2.3 Proximity Relationship

The proximity is defined as the nearness or closeness between domains of tuples in a relationship. This proximity can be identified by using the Euclidean distance method. The Euclidean distance gives the closeness of one attribute to another. This article proposed the proximity relation based method for finding the dependencies between attributes in a relational model and a new algorithm for extracting dependencies based on conformance and Mamdani technique [6].

Here we are going to identify the existing dependencies between various tuples in Fuzzy Relational Data Base (FRDB). The FRDB model also has redundancy like traditional model. So we have to reduce that by using normalization techniques. It offers the well designed database which is not affected by update anomalies. It is done by finding closeness and closure of attributes [7]. Such redundancy is reduced in databases by using closeness between attributes in a relation. The fuzzy database is object oriented in nature and similarity degree applies to that to represent the fuzziness. It depends on partition and equality classes. Fuzzy Logic is taking care of the process of assigning classes. Fuzzy Rule is developed by doctors for updating the symptoms and the details of class in the system [8]. FRDB models are classified as Similarity and Possibility model. Avoiding imperfect information in real world applications may lead to loss of necessary information. This can be rectified by fuzzy dependencies which are understood using the conformance method [9]. This gives the hidden and useful information about the data set. This information is very useful to make better decision. Handling of inaccuracies is an ordeal job and needs more time and effort, Managing and representing unnecessary and inaccurate information is powerfully done by means of fuzzy theory and sets. If the database managers know

the dependencies of attribute it should help in reducing redundant entries in a relation and also for updating anomalies.These can be done by means of nearness or proximity between attributes [10]. It gives the method for dealing with the unnecessary data in a database. It is a very tough problem and it is solved by using extended proximity relation. Here the formula of proximity relation should be modified slightly and used to solve the problem thus mentioned [11].

The following attributes are age, sex, cp, bp, cholesterol, fbs, ecg, thalach.
Attribute domains of relations:
D (age) = {middle, old}
D (sex) = {male, female}
D (cp)= {typical angina, typical angina,non-anginal pain, asymptomatic}
D (bp) = {middle, high}
D (cholesterol) = {normal, high}
D (fbs) = {true, false}
D (ecg) = {normal, left ventricular hypertrophy}
D (thalach) = {normal, high}

The steps to calculate the proximity are given below.

1. Take the values (table-1) of instances for the specified attribute. Set remaining instance space as 0.
2. Subtract the value of one instance with others.
3. Then take the square of each value obtained from subtraction.
4. Add the square values. Divide this sum by the total number of tuples.
5. Then take the square root for resultant value.Divide this value by condition range limits (table-3).
6. Do, for all attributes.

The proximity relation technique is more superior to other techniques in the literature. This could overcome the difficulties that were in previous approaches, for example, a similarity relation.Here we can satisfy the reflexive property, binary relationships of fuzzy and symmetric nature of the data sets. This is the very strong and sound definition of proximity. The algorithm gives better result than similarity relations [12].table-4 describes the closeness measure of age attribute denoted by S_{Age}. This tuple is taken as the critical attribute and its conformance C (Age) assumed to be 0.70. The table-5 describes the Closeness Measure of Sex attribute denoted byS_{Sex}. The table-6 defines the Closeness Measure of Chest Pain Type tuple denoted by $S_{ChestPain}$.

Table-4: Closeness Measure of Age

S_{Age}	**Middle**	**Old**
Middle	1.00	0.70
Old		1.00

Table-5: Closeness Measure of Sex

S_{Sex}	Male	Female
Male	1.00	0.71
Female		1.00

Table-6: Closeness Measure of Chest Pain Type

S_{CP}	Asymptomatic	Typical Angina	Non-Anginal Pain	Atypical Angina
Asymptomatic	1.00	0.86	0.93	0.40
Typical Angina		1.00	0.91	0.76
Non-Anginal-Pain			1.00	0.56
Atypical Angina				1.00

The table-7 defines the Closeness Measure of Blood Pressure tuple denoted by $S_{BloodPressure}$. The content depicted in table-8 the Closeness Measure of Cholesterol attribute denoted by$S_{Cholestrol}$. The table-9 depicts the Closeness Measure of Fasting Blood Sugar (FBS) denoted byS_{FBS}. The table-10 describes the Closeness Measure of Electro Cardio Gram (ECG) tuple and is denoted by S_{ECG}. The table-11 denotes the Closeness Measure of Thalach tuple and it is denoted by $S_{Thalach}$.

Table-7: Closeness Measure of Blood Pressure

$S_{BloodPressure}$	Normal	High
Normal	1.00	0.89
High		1.00

Table-8: Closeness Measure of Cholesterol

$S_{Cholesterol}$	Normal	High
Normal	1.00	0.92
High		1.00

Table-9: Closeness Measure of Fasting Blood sugar

S_{FBS}	Yes	No
Yes	1.00	0.80
No		1.00

Table-10: Closeness Measure of Electro Cardio Gram(ECG)

S_{ECG}	Normal	Left Ventricular Hypertrophy
Normal	1.00	0.83
Left Ventricular Hypertrophy		1.00

Table-11: Closeness Measure of Thalach

$S_{Thalach}$	Normal	High
Normal	1.00	0.77
High		1.00

Numerical Example to Calculate the Proximity Relation for a fuzzy attribute from the cardiac data set is given below:

Proximity Relation for Blood Pressure (BP):

S_{BP}= (√((0-145)^2+(0-160)^2+(120-0)^2+(130-0)^2+(130-0)^2+(120-0)^2+(0-140)^2+(120-0)^2+(130-0)^2+(0-140)^2)/total number of tuples)/condition range limits =0.89

Where, number of tupels =10 and condition range limits= 150 (table-2 and table-3 respectively).

2.4 Conformance

Conformance defined as the rate of similarity between attributes. It ranges from 0 to 1. The statement of possibility distribution and probability distribution is given by using closeness relation that is proximity. How much closeness exists between two attributes of relation is known as similarity of a relation i.e. Conformance. These can be used in medical diagnosis, recognition of speech & patterns, etc., [13]. The DB can be either of an application based type or Object Oriented (OO) type. Depending on the context set is selected. Both the types have inaccuracies to be dealt. Handling of ambiguous information on real world can be achieved by fuzzy methods. The OO based DB can be used to manage the complex objects and find the relationships which are unnecessary. This can be done by means of similarity degree method [14]. The similarity degree and nearness of a relation go hand in hand. Relationship occurs in between attributes in a data set which can be identified by applying conformance measure. This conformance is combined with the closeness to yield better results [15]. The supporting agent to any of the existing conformance approaches is the Tmin value. In FFDs t-norms have been used in the similarity definition in the relation and degree of fuzziness assigned to it. Also a finite logic has been proposed to delete repeated data [16].

The values for the conformance of the attributes of the data set are tabulated in table-11.Conformance of B_l for any two tuples (t_m,t_n) of a domain in a relation r is given in Equation (1),

$$C(B_l[t_m, t_n]) = \wedge\{\bigwedge_{x \epsilon d_m}\{\vee_{y \epsilon d_n}\{s(x,y)\}\}\},$$

$$\wedge_{x \epsilon d_n}\{\vee_{y \epsilon d_m}\{s(x,y)\}\}\} \quad (1)$$

The condition for the FFD $X \overset{\theta}{\rightarrow} Y$ to be satisfied is in Equation (2).

$$C(Y\ [t_m, t_n]) \geq \min(\theta, C(X\ [t_m, t_n])) \quad (2)$$

Where $C(B_l[t_m, t_n])$ −Conformance of a attribute B_1, d_m-Attribute B_1 value for tuple t_m, d_n-Attribute B_1 value for tuples t_n , s (x, y) –

similarity relation of values x, y , θ-Strength of linguistic parameter (ranges from 0 to 1), ∨-Maximum value selection, ∧-Minimum.

$\boldsymbol{\theta_k}$-Dependency Strength of Fuzzy variable with value equal to 0.7 [1]. It is an important parameter used in table-12 that depicts the Conformance value for Cardiac Dataset that denotes the relational understanding between every instance of an attribute to every other instance of an attribute.

Table-12: Conformance value for Cardiac Dataset

$\boldsymbol{\theta_k}$	C(A[$\boldsymbol{t_i}, \boldsymbol{t_j}$])	C(S[$\boldsymbol{t_i}, \boldsymbol{t_j}$])	C(Cp[$\boldsymbol{t_i}, \boldsymbol{t_j}$])	C(B[$\boldsymbol{t_i}, \boldsymbol{t_j}$])	C(C[$\boldsymbol{t_i}, \boldsymbol{t_j}$])	C(F[$\boldsymbol{t_i}, \boldsymbol{t_j}$])	C(E[$\boldsymbol{t_i}, \boldsymbol{t_j}$])	C(T[$\boldsymbol{t_i}, \boldsymbol{t_j}$])
$\mathbf{t_{1,2}}$	0.70	1.00	0.86	1.00	0.92	0.80	1.00	1.00
$\mathbf{t_{1,3}}$	0.70	1.00	0.86	0.89	1.00	0.80	1.00	1.00
$\mathbf{t_{1,4}}$	0.70	1.00	0.91	0.89	0.92	0.80	0.83	0.77
$\mathbf{t_{1,5}}$	0.70	0.71	0.76	0.89	1.00	0.80	1.00	0.77
$\mathbf{t_{1,6}}$	0.70	1.00	0.76	0.89	1.00	0.80	0.83	0.77
$\mathbf{t_{1,7}}$	0.70	0.71	0.86	1.00	0.92	0.80	1.00	0.77
$\mathbf{t_{1,8}}$	0.70	0.71	0.86	0.89	0.92	0.80	0.83	0.77
$\mathbf{t_{1,9}}$	0.70	1.00	0.86	0.89	0.92	0.80	1.00	1.00
$\mathbf{t_{1,10}}$	0.70	1.00	0.86	1.00	1.00	1.00	0.83	1.00
$\mathbf{t_{2,3}}$	0.70	1.00	1.00	0.89	0.92	1.00	1.00	1.00
$\mathbf{t_{2,4}}$	0.70	1.00	0.93	0.89	1.00	1.00	0.83	0.77
$\mathbf{t_{2,5}}$	0.70	0.71	0.40	0.89	0.92	1.00	1.00	0.77
$\mathbf{t_{2,6}}$	0.70	1.00	0.40	0.89	0.92	1.00	0.83	0.77
$\mathbf{t_{2,7}}$	0.70	0.71	1.00	1.00	1.00	1.00	1.00	0.77
$\mathbf{t_{2,8}}$	0.70	0.71	1.00	0.89	1.00	1.00	0.83	0.77
$\mathbf{t_{2,9}}$	0.70	1.00	1.00	0.89	1.00	1.00	1.00	1.00
$\mathbf{t_{2,10}}$	0.70	1.00	1.00	1.00	0.92	0.80	1.00	1.00
$\mathbf{t_{3,4}}$	0.70	1.00	0.93	1.00	0.92	1.00	0.83	0.77
$\mathbf{t_{3,5}}$	0.70	0.71	0.40	1.00	1.00	1.00	1.00	0.77
$\mathbf{t_{3,6}}$	0.70	1.00	0.40	1.00	1.00	1.00	0.83	0.77
$\mathbf{t_{3,7}}$	0.70	0.71	1.00	0.89	0.92	1.00	1.00	0.77
$\mathbf{t_{3,8}}$	0.70	0.71	1.00	1.00	0.92	1.00	0.83	0.77
$\mathbf{t_{3,9}}$	0.70	1.00	1.00	1.00	0.92	1.00	1.00	1.00
$\mathbf{t_{3,10}}$	0.70	1.00	1.00	0.89	1.00	0.80	1.00	1.00
$\mathbf{t_{4,5}}$	0.70	0.71	0.56	1.00	0.92	1.00	0.83	1.00
$\mathbf{t_{4,6}}$	0.70	1.00	0.56	1..00	0.92	1.00	1.00	1.00
$\mathbf{t_{4,7}}$	0.70	0.71	0.93	0.89	1.00	1.00	0.83	1.00
$\mathbf{t_{4,8}}$	0.70	0.71	0.93	1.00	1.00	1.00	1.00	1.00
$\mathbf{t_{4,9}}$	0.70	1.00	0.93	1.00	1.00	1.00	0.83	0.77
$\mathbf{t_{4,10}}$	0.70	1.00	0.93	0.89	0.92	0.80	0.83	0.77

$t_{5,6}$	0.70	0.71	1.00	1.00	1.00	1.00	0.83	1.00
$t_{5,7}$	0.70	1.00	0.40	0.89	0.92	1.00	1.00	1.00
$t_{5,8}$	0.70	1.00	0.40	1.00	0.92	1.00	0.83	1.00
$t_{5,9}$	0.70	0.71	0.40	1.00	0.92	1.00	1.00	0.77
$t_{5,10}$	0.70	0.71	0.40	0.89	1.00	0.80	1.00	0.77
$t_{6,7}$	0.70	0.71	0.40	0.89	0.92	1.00	0.83	1.00
$t_{6,8}$	0.70	0.71	0.40	1.00	0.92	1.00	1.00	1.00
$t_{6,9}$	0.70	1.00	0.40	1.00	0.92	1.00	0.83	0.77
$t_{6,10}$	0.70	1.00	0.40	0.89	1.00	0.80	0.83	0.77
$t_{7,8}$	0.70	1.00	1.00	0.89	1.00	1.00	0.83	1.00
$t_{7,9}$	0.70	0.71	1.00	0.89	1.00	1.00	1.00	0.77
$t_{7,10}$	0.70	0.71	1.00	1.00	0.92	0.80	1.00	0.77
$t_{8,9}$	0.70	0.71	1.00	1.00	1.00	1.00	0.83	0.77
$t_{8,10}$	0.70	0.71	1.00	0.89	0.92	0.80	0.83	0.77
$t_{9,10}$	0.70	1.00	1.00	0.89	0.92	0.80	1.00	1.00

C ($A[t_i,t_j]$) -Conformance of Attribute Age, **C($S[t_i,t_j]$)** -Conformance of Attribute Sex, **C($Cp[t_i,t_j]$)** - Conformance of Attribute Chest Pain Type, **C($B[t_i,t_j]$)** - Conformance of Attribute Blood Pressure, **C($Ch[t_i,t_j]$)** -Conformance of Attribute Cholesterol, **C($F[t_i,t_j]$)** -Conformance of Attribute Fasting Blood Sugar, **C($E[t_i,t_j]$)** - Conformance of Attribute ECG, and **C($T[t_i,t_j]$)** -Conformance of AttributeThalach.

Numerical example for conformance value calculation for an attribute is given below: Conformance of attribute Blood Pressure for instances t_1 and t_3 is calculated using Equation (1).

C(Blood Pressure BP[t_1,t_3]) = min{min{max{s (High,Normal)}},min{max{s(Normal,High)}}}= min{min{max{s(0.89)}},min{max{s(0.89)}}} = min{min{0.89},min{0.89}}= min{0.89,0.89} = 0.89

3. Multi Domain Fuzzy Cluster

The MDFC (Multi Domain Fuzzy Cluster) is an extension of classical fuzzy technique. As the name suggests, this fuzzy cluster applies to a heterogeneous domain of inputs considered as a matrix structure. The fuzzy clusters in many literatures have been applied to group the given set of based on class instances for invariably some type of application taken under consideration. In this proposed approach, since the problem adheres to the understanding of the relationship between tuples of the data set, a new dimension of clustering is proposed based on the 1 to N correspondence.

Generally, real time DB are fuzzy in nature, they tend to have multiple values. This article describes a heuristic algorithm for dealing with such multi valued instances. They prove better than the existing techniques[17].

The clustering method is best suited for dealing with raw data. Data upon being grouped gets transformed into meaningful information. This article explores the time series data by clustering them based on Dynamic Time Warping distance [18]. Fuzzy DBs (Data Bases) can’t accept simple SQL they need a new form of SQL to be accessed. In this article Fuzzy SQL (Structured Query Language) is introduced based

on EFCM(Extended Fuzzy C Means) approach supported by GK(Gustafson Kessel) algorithm [19]. Most of the time Fuzzy clustering is applied in a hybrid fashion. This article proposed a hybrid fuzzy model that uses Support Vector Machines, Genetic Algorithm and Fuzzy C-means to obtain better results [20].

The pseudo code for MDFC approach is given below.

1. Set input as X. Here X is a matrix (Fuzzy Dependency Relationship) of order m×n.
2. Set a centroid C. Here C is a matrix (Classical Dependency Relationship) of order 3×3.
3. Now take X[i] (instance of FDR matrix) and perform Multi Domain Fuzzy Clustering with C.
4. Continue step 3 till all instances of X matrix is covered.
5. Store each value of Multi Domain Fuzzy Clustering(X[i],C) in S[i].
6. Take set S and compare each instance in it for following condition

- If (S[i]>=0 and S[i]<=0.5) then club it in Densely Cluster(DRC)
- If (S[i]>0.5 and S[i]<=0.7) then club it in Loosely Densely Cluster(LDC)
- If(S[i]>0.7 and S[i]<=1) then club it in Strongly Densely Cluster(SDC).

7. The DC is a preferable region of the cluster and the data set is in safe zone, SDC is strongly recommended region of the cluster and is also in the safe zone and LDC is in look after region of the cluster and is the critical zone of treatment.
8. The C matrix values correspond to the alpha cut values in fuzzy sets, namely α=0, 0.5, 1
9. A simple count wise comparison is done to predict the movement of data sets into the bags.

The mathematical model of MDFC is discussed below.

$$Euclidian\ Distance\ d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

Where x_i is instance of X(input) , y_i is an instance of C (centroid) , n is the column size in matrix (X and C)

$$Total\ Distance\ Td = \sum_{i=1}^{n} 1/di \quad (4)$$

where d_i is the Euclidian distance of every instance , n is the total number of Euclidian distance

$$Fuzzy\ Cluster\ Fc = (\frac{d_i}{Td})^{b-1} \quad (5)$$

Where i=1, 2, 3… n; d_i -Euclidian distance of each instance , b is assumed to be 2 as per literature.

Table-13: Tmin of cardiac data set

ATTRIBUTES	Age	Sex	CP	BP	Cholesterol	FBS	ECG	Thalach
Age	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
Sex	0.70	0.71	0.75	0.71	0.75	0.75	0.75	0.75
CP	0.70	0.71	0.75	0.76	0.76	0.75	0.76	0.76
BP	0.70	0.71	0.76	0.89	0.89	0.80	0.83	0.77
Cholesterol	0.70	0.71	0.76	0.89	0.92	0.80	0.83	0.77
FBS	0.70	0.71	0.76	0.80	0.80	0.80	0.80	0.77
ECG	0.70	0.71	0.76	0.83	0.83	0.80	0.83	0.77
Thalach	0.70	0.71	0.77	0.77	0.77	0.77	0.77	0.77

Table-14: Tmin of user defined data set

ATTRIBUTE	A1	A2	A3	A4	A5	A6	A7	A8
A1	0.2	0.2	0.2	0.1	0.2	0.1	0.1	0.2
A2	0.2	0.1	0.1	0.2	0.1	0.1	0.2	0.2
A3	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.2
A4	0.1	0.2	0.2	0.1	0.2	0.2	0.2	0.1
A5	0.2	0.1	0.1	0.1	0.1	0.2	0.1	0.1
A6	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.2
A7	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.2
A8	0.1	0.2	0.2	0.1	0.2	0.1	0.1	0.2

Numerical calculation for MDFC approach is given below :
Consider two matrix X as input and C as centroids

$$X = \begin{bmatrix} 0.70 & 0.70 & 0.70 & 0.70 & 0.70 & 0.70 & 0.70 & 0.70 \\ 0.70 & 0.71 & 0.75 & 0.71 & 0.75 & 0.75 & 0.75 & 0.75 \\ 0.70 & 0.71 & 0.75 & 0.76 & 0.76 & 0.75 & 0.76 & 0.76 \\ 0.70 & 0.71 & 0.76 & 0.89 & 0.89 & 0.80 & 0.83 & 0.77 \\ 0.70 & 0.71 & 0.76 & 0.89 & 0.92 & 0.80 & 0.83 & 0.77 \\ 0.70 & 0.71 & 0.76 & 0.80 & 0.80 & 0.80 & 0.80 & 0.77 \\ 0.70 & 0.71 & 0.76 & 0.83 & 0.83 & 0.80 & 0.83 & 0.77 \\ 0.70 & 0.71 & 0.77 & 0.77 & 0.77 & 0.77 & 0.77 & 0.77 \end{bmatrix}$$

$$C = \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix}$$

Upon applying, Euclidian distance formula in Equation (3) is used to X[1] and C[i] where i=1,2,3

d(X[1] C[1]) = 1.98

d(X[1] C[2]) =0.56

d(X[1] C[3]) =0.84

Total Euclidian distance formula from Equation (4) is used as [1/d(X[1] C[1]) + 1/d(X[1] C[2]) + 1/d(X[1] C[3])]

Td =3.48

Fuzzy Cluster formula from Equation (5) is used as Fc = d[i]/Td
Where i= X [1] C [1 [, X [1] C[2], X[1] C[3]

Fc1 =0.57

Fc2 =0.14

Fc3=0.21

Now find the greatest of value which is Fc1 and so this input instance is Loosely Densely Cluster (LDC). Likewise repeat the steps for another set of inputs also and by using count wise comparison technique compute the final data set clustering region.

3. Results and Discussion

The quality of any work is admired by its outcome. In this context attempt are made to view data sets and their implicit corresponding relationship in a new angular domain. This result shall contribute to the efficacy of data set and also improve on the thought process of users who collect and present data for further colorful analysis using various methodologies available in the literature. The MDFC approach would club data into three different bags of cluster, namely DC, LDC and SDC; which would help end users appreciate the strength of the set thus collected. To provoke this approach to greater standards, it is applied over user defined as well as real time data sets and blessed results are observed. The clustered pattern for the respective fuzzy relational data set is represented graphically as shown in Fig.2 and Fig.3 which is bar graphs representing real time data set which is cardiac disease set and user defined data set respectively for enhanced understanding of the relationship by which the facts are governed.

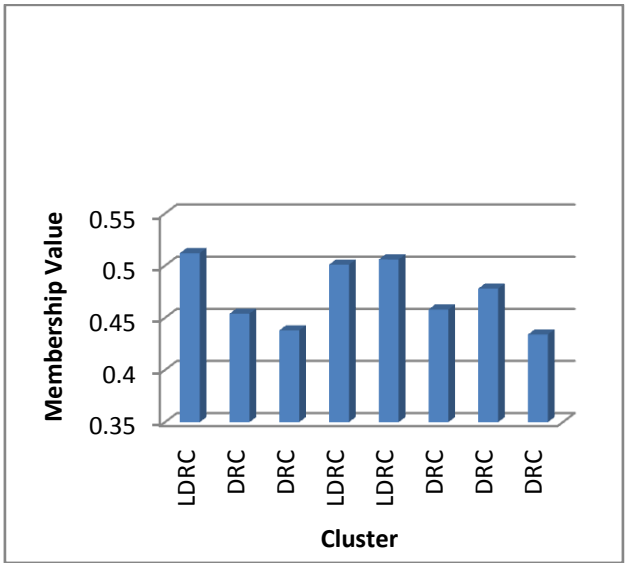

Fig.2 : Bar Graph for Cardiac Data Set

The given real time fuzzy data set of cardiac disease whose Tmin values are tabulated in table-13, Tmin of the cardiac data set is formed to a matrix and applied MDFC to it to get itself clustered to the preferable safe zone of Densely Cluster(DC) when solved using a heuristic program (Fig.2). It is observed from Figure 2 that out of eight membership values; five corresponds to the DC and three to LDC. The user defined Tmin matrix that is formed using the able-14, Tmin of user defined data set is clustered into the critical region which is Loosely Densely Cluster (Fig.3). It is observed from Fig. 3 that all the eight values correspond to the LDC region.

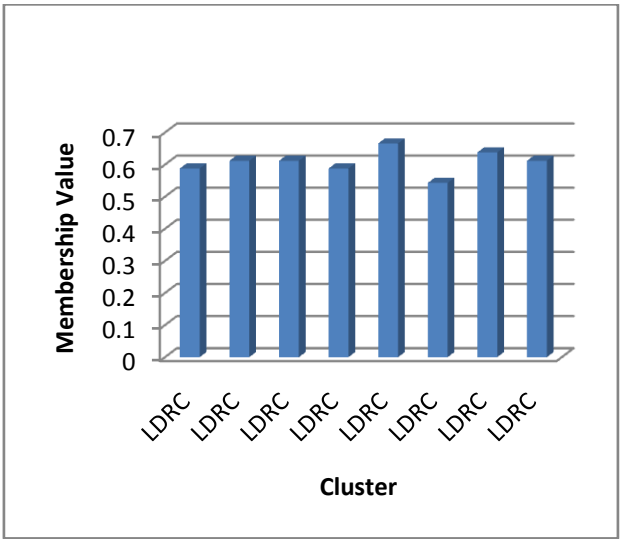

Fig .3: Bar Graph for User Defined Data Set

4. Conclusion

The understanding of an implicit relationship of any data set has remain in to chat level for a very long period of time. It is only now that people have identified the essential need for such relationship to be analyzed and to be determined. This pre-analyzer concept would eventually become a hallmark for data sets as the approach concentrates on the evaluation of the input as a whole. The advantage of MDFC approach is that it can save a lot of time and effort by giving end users a pre-requisite of the data that is collected and formulated as a data set. It can also reduce the cost of failures thus caused upon the implementation of methodologies available in the literature over the data set. It gives a trade-off between cost times quality called Cost Time Quality Trade-off Problem(CTQTP). The drawback of MFDC is the handler conditions to bag sets to different clusters. The handler condition concentrating on the region of critical section needs more effort and study in future for enhancing the data sets. In the future attempts, methods are to be proposed for managing the handler conditions and develop approaches to bring LDC sets to either of DC or SDC sets.

References

[1] Graves, D., Noppen, J., and Pedrycz, W.,2012, "Clustering with proximity knowledge and relational knowledge," Pattern Recognition, pp. 2633–2644.

[2] Chakraborty, S., 2012, "Codd's Relational Data Model and Fuzzy Logic : A Practical Approach to Find the Computer Solution," International Journal of Advanced Technology & Engineering Research, Vol. 2, pp. 21-27.

[3] García-Lapresta, J. L., and Pérez-Román, D., 2015, "Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering," Applied Soft Computing, pp. 2-9.

[4] Yu, J., and Liu, C.,2014, "Lossless Join Decomposition for Extended Possibility-Based Fuzzy Relational Databases," Journal of Applied Mathematics, Vol. 2014, pp. 1-9.

[5] Kupka, J., and Tomanovà, I.,2012, "Dependencies among attributes given by fuzzy confirmation measures," Expert Systems with Applications, 39, pp. 7591-7599.

[6] Vucetic, M., Hudec, M., and Vujošević, M.,2013, " A new method for computing fuzzy functional dependencies in relational database systems," Expert Systems with Applications, pp. 2738–2745.

[7] Mishra, J., and Ghosh, S.,2012," Normalization in a fuzzy relational database model," International Journal of Computer Engineering and Technology, Vol. 3, pp. 506-517.

[8] Dutta, S., Sahoo, L., and Dwibedy, D.,2013, "An Equivalence Relation to Reduce Data Redundancy Based on Fuzzy Object Oriented Database System," International Journal of Soft Computing and Engineering, Vol. 3, pp. 417-421.

[9] Anupriya, and Rishi, R.,2013,"Review Fuzzy Logical Database Models," IOSR Journal of Computer Engineering, Vol. 8, pp 24-30.

[10] Vučetić, M.,2011," Functional dependencies analyse in fuzzy relational database models," Journal of Information Technology and Applications, Vol. 1, no. 1, pp. 90-104.

[11] Rajaei, A., Dastjerdi, A. B., and Aghaee, N, G.,2011, "An extension of semantic proximity for Fuzzy multivalued dependencies in fuzzy Relational database," International Journal of Database Management Systems, Vol.3, No.3, pp 157-169.

[12] Julián-Iranzo, P., and Rubio-Manzano, C.,2015," Proximity-based unification theory," Fuzzy Sets and Systems, pp. 21-43.

[13] Shukla, P. K..,Darbari,M.,Singh,V.K., and Tripathi,S.P.,2011,"A Survey of Fuzzy Techniques in Object Oriented Databases," International Journal of Scientific & Engineering Research, Vol. 2, pp. 1-11.

[14] Zadeh, L. A.,2014," A note on similarity-based definitions of possibility and probability," Information Sciences, pp. 334–336.

[15] Kupka, J., and Tomanová, I.,2011,"Some dependencies among attributes given by fuzzy confirmation measures.," EUSFLAT-LFA, pp. 498-505.

[16] Cordero,P., Enciso,M., Mora,A., and de Guzmnán,I.P.,2010," A complete logic for fuzzy functional dependencies over t-norms," Congreso Español Sobre Tecnologías y Lógica Fuzzy, 205-210.

[17] Angryk, R. A., and Czerniak, J.,2010, "Heuristic algorithm for interpretation of multi-valued attributes in similarity-based fuzzy relational databases," International Journal of Approximate Reasoning, pp. 895-911.

[18] Izakian, H., Pedrycz, W., and Jamal, I.,2015," Fuzzy clustering of time series data using dynamic time warping distance," Engineering Applications of Artificial Intelligence, pp. 235-244.

[19] Jain, N., and Shukla, S.,2012,"Fuzzy Databases Using Extended Fuzzy C-Means Clustering," International Journal of Engineering Research and Applications, Vol. 2, pp. 1444-1451.

[20] Aydilek, I. B., and Arslan, A., 2013, " A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," Information Sciences, pp. 25-35.