Fuzzy C-Means based Inference Mechanism for Association Rule Mining: A Clinical Data Mining Approach

Kapil Chaturvedi
Dept. of Computer Application, RGPV,
Bhopal, MP, India

Dr. Ravindra Patel
Dept. of Computer Application, RGPV
Bhopal, MP, India

Dr. D.K. Swami
Faculty of Engineering,
VNS Group of Inst.
Bhopal, MP, India

Abstract—**Association rule mining has wide variety of research in the field of data mining, many of association rule mining approaches are well investigated in literature, but the major issue with ARM is, huge number of frequent patterns cannot produce direct knowledge or factual knowledge, hence to find factual knowledge and to discover inference, we propose a novel approach AFIRM in this paper followed by two step procedure, first is to discover frequent pattern by Appling ARM algorithm and second is to discover inference by adopting the concept of Fuzzy c-means clustering, for performance analysis, we apply this approach on a clinical dataset (contained symptoms information of patients) and we got highly effected disease in a couple of months or in a session as hidden knowledge or inference.**

Keywords—Association Rule Mining; Fuzzy Inference System; Clinical Data Mining; Preprocessing; Fuzzy clusters

I. INTRODUCTION

Association rule mining (ARM) is the well-researched data mining technique [10, 15]. To find the frequent relation or association between items from market basket analysis perspective, which uses a rule based knowledge representation to find the relationship or causal dependencies between objects, things, attributes, outcomes, symptoms, occurrences, etc. It was first introduced in 1993 as the AIS algorithm [2] then in 1994 R. Agrawal and R. Srikant provided a candidate generation based technique formally known as Apriori algorithm [1] to generate rules, it is a streamlined version of AIS algorithm, it outperforms when support count is high and a number of items are less. The second approach for ARM is Frequent Pattern growth formally known as FP- Growth algorithm [11] proposed by J. Han, J. Pei and Y. in 2000.

But in present era we realize that association rule mining is a strong approach not only for market basket analysis rather than it plays vital role in various data analysis fields like stock market analysis, DNA pattern recognition, web data mining and also in clinical data mining, but the limitation of Association rule mining is it produces huge numbers of frequent patterns as per predefined thresholds which is insufficient to draw a conclusion.

The aim of this research work is to design and develop an inference mechanism for association rule mining, in order to discover abstract knowledge from the huge number of frequent patterns. Under this research, we proposed new algorithms, AFIRM to achieve the required goal. The major objectives of this research can be summarized in following points.

- To investigate the literature and problem identification.

- To choose a suitable approach for association rule mining for finding frequent patterns.

- To design an inference mechanism framework for association rule mining.

- To develop an inference mechanism for association rule mining in order to discover abstract/inference knowledge from frequent patterns.

- To analyze the result to meet abstract inference knowledge.

In this paper, we proposed a Fuzzy C-means (FCM) based inference system in order to discover inference knowledge, and the rest of the paper is organized as follows: section 2 elaborates the detail about fuzzy inference mechanism, section 3 give a comparison between hard and soft clustering approaches, section 4 introduces a literature review with basic terminology and previous research with related work already done in respective area, section 5 presented a comparative study between Apriori and FP-growth algorithm in order to choose most suitable ARM approach, section 6 proposed new approach along with illustrative example, simulation and result is presented in section 7, finally concluding remark is given in section 8.

II. INFERENCE MECHANISM

An inference engine is developed for an expert system consisting of inference mechanism as well as a control strategy. The term inference refers to the process of searching through the knowledge base and deriving new knowledge [18].

It involves formal reasoning by matching and unification, similar to the one performed by human experts to solve problems in a specific area of knowledge. An inference rule may be defined as a statement that has two parts an if clause and a then clause. It can be understood more clearly by adopting following example.

Rule 1: If Symptoms are headache, sneezing, running_nose then the patient have cold

Rule 2: if Symptoms are fever, cough and running nose, then patients have measles

III. Soft V/S Hard Clustering

Fuzzy c-means clustering algorithm is a little bit different from traditional clustering approaches due to its fuzzy nature and also because of the capacity of handling delicate data. It first introduced by Dunn [8] and then improved by Bezdek [3]. Traditional clustering algorithms (K-means [19], K-medoids [17]) are also known as hard clustering techniques because it divides the data in distinct clusters where each data element belongs to the exactly one group as shown in Figure-1 and 2.
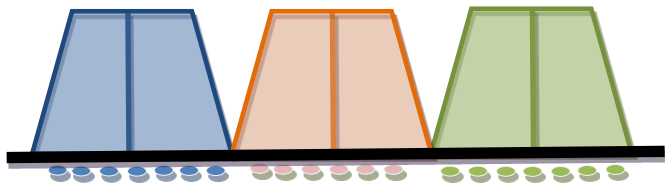

Fig. 1. Partitioning of data on the basis of similarity in traditional or Hard clustering approaches
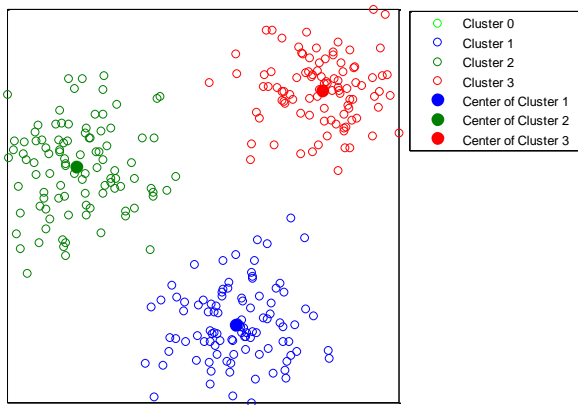

Fig. 2. Data clustering in traditional or Hard clustering approaches

Whereas fuzzy, c-means clustering or soft clustering algorithm provides an extended capacity of data categorization [7] where data elements can fit into more than one cluster and is associated with each object as shown Figure-3 & 4. This is a set of membership levels that specify the strength of the association between the data objects and a particular cluster. Fuzzy clustering is a process of assigning such membership levels and then using them to assign data objects to one or more clusters.
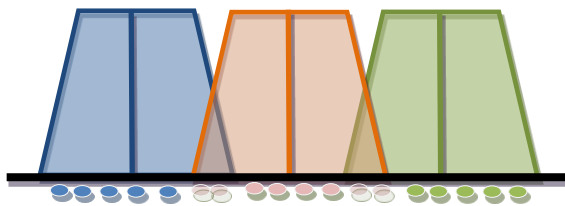

Fig. 3. Partitioning of data on the basis of degree of membership in fuzzy clustering approach
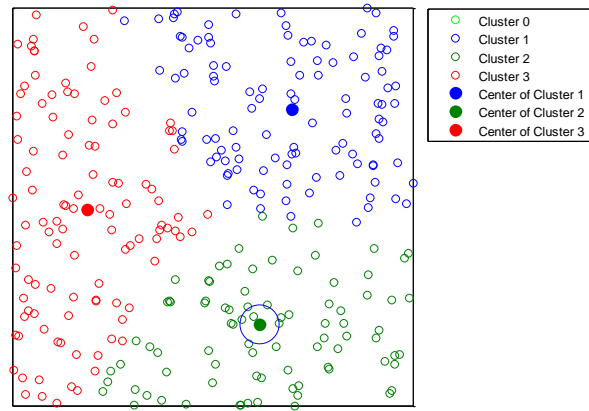

Fig. 4. Data clustering by fuzzy, c-means clustering approach

It is a strong approach from the data analysis perspective because it provides the way of organizing data into multiple predefined clusters or groups, based on their similarity among the classes from which it belongs to. But the additional extensibility of this approach is a common itemset that may refer to multiple clusters based on their degree of membership or similarity. Similarity refers to the mathematical calculation of similarity in the general distance calculation between objects using some well-defined distance measures.

A. Algorithm: Fuzzy C-Means (FCM)

Fuzzy c-means clustering algorithm is most suitable approach for implementing a fuzzy inference system, in this regard; we adopt the FCM algorithm to find factual knowledge hide in frequent patterns generated by ARM algorithm.

Input: D dimensional matrix, Number of Clusters – C

Output: Cluster Center Cq, Membership matrix U=[Uij]

Step-1: Let D dimensional matrix U with p data points represented by U=Upq, where initialize U=U (0)

Step-2: Assign the pre-defined number of clusters C, where 2≤C≤n

Step-3: At k-step calculate the center vector Cq, Where the cluster center C_q can calculate as follows:

$$C_q = \frac{\sum_{p=1}^{N} U_{pq}^m - X_p}{\sum_{r=1}^{C} U_{pq}^m}$$

Step-4: Calculate the degree of membership for pth data point in cluster q can be calculated as follows.

$$U_{pq} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|X_p - C_q\|}{\|X_p - C_r\|} \right)^{\frac{2}{m-1}}}$$

Step-5: Update Ur, Ur+1 (as STEP-4) at each step

Step-6: If $\left| U_{pq}^{r+1} - U_{pq}^{r} \right| < \varepsilon$, Then STOP

(Where, ε is the termination decisive factor or pre specified termination criteria between 0 and 1)

Step-7: Else return to Step-3

IV. RELATED WORK

An inference mechanism framework for association rule mining proposed in [4] presents a theoretical and numerical study on association rule based inference mechanism for discovering factual knowledge from a clinical dataset in order to discover highly effected disease in a particular session by proposing an algorithm AIRM while this paper is presented a modified AIRM algorithm as AFIRM algorithm further describe in the next section.

In this manner to review the literature and to study on related work we investigate some good research approaches in the respective field, we also review the approaches for clinical data mining and fuzzy inference based approach.

A. Inference approaches for ARM

In [9] Ronald Fagin et al. presents a brief overview of inference rules, they also gave a brief discussion on the applicability of inference in various areas like inference in propositional logic, non-standard propositional logic, propositional modal logic and inference in first order logic, etc, X. Wanga et. al[23] proposed a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data, for this they made use of the cluster information generated by a self-organizing map for pattern analysis and a fuzzy inference system to capture the chaotic trend to provide short-term (hourly) and long-term (daily) Web traffic trend predictions. Yang et al. [24] proposed an approach generic rule-base inference methodology using the evidential reasoning (RIMER) in this they proposed a new knowledge representation scheme in a rule base by analyzing existing knowledge base structure using a belief structure. Similarly R. Chow et. al [6] provides an ARM based Inference Mechanism, in this paper, they propose a refined and practical model of inference detection using a reference corpus. This model is inspired by association rule mining: inferences are based on word co-occurrences. Using the model and taking the Web as the reference corpus, that model also includes the important case of private corpora, to model inference detection in enterprise settings in which there is a large private document repository. They found inferences in private corpora by using analogues of his Web-mining algorithms, relying on an index for the corpus rather than a Web search engine. Tree-based mining approach for discovering patterns of human interaction in meetings [25] presented by Zhiwen Yu et. al for mining, human interactions for accessing and understanding meeting content for this they proposed a tree-based mining method for discovering frequent patterns of human interaction

in meeting discussions. As per this study the mining results would be useful for summarization, indexing, and comparison of meeting records and interpretation of human interaction in meetings.

B. ARM approaches for clinical data mining

Data mining is more popular in the field of medical science due to its high applicability and analysis ability in medical and clinical data mining here we review some previous approaches related to medical field.

S. Venus et al. [21] proposed a rule based backward chaining inference engine which is an Arabic expert system based approach on natural language for diagnosing diseases, similarly Yanqing Ji et. al presents a potential causal association mining algorithm [16] for screening, adverse drug reactions in post marketing surveillance and proposed a novel data mining approach to signaling potential ADRs (Adverse drug reaction) from electronic health databases they also introduced potential causal association rules (PCARs) to represent the potential causal relationship between a drug and ICD-9 [13] coded signs or symptoms representing potential ADRs. Due to the infrequent nature of ADRs, the existing frequency-based data mining methods cannot effectively discover PCARs. They introduce a new interesting measure, potential causal leverage, to quantify the degree of association of a potential causal association rule (PCAR) similarly paper [20, 22] proposed practical and applied fuzzy logic based approaches for medical diagnosis, whereas paper [5] presents a novel data mining approach to generate adverse drug events detection rules. The main objective of this paper is to automatically detect cases of ADEs (adverse drug events) by data mining. They used decision trees and association rules to discover ADE detection rules, with respect to time constraints. The rules are then filtered, validated, and reorganized by a committee of experts. The rules are described in a rule repository, and several statistics are automatically computed in every medical department, such as the confidence, relative risk, and median delay of outcome appearance.

V. PERFORMANCE ANALYSIS OF APRIORI & FP-GROWTH ALGORITHMS

In this section, performance analysis has been carried out in order to analyze the efficiency of Apriori and FP-Growth algorithm moreover choose the most appropriate ARM algorithm for finding frequent patterns, for this purpose we apply the both algorithms on T10I4D100K [12] and pima D38.N768.C2 [14] datasets, and measure time efficiency on different support counts, as shown in figure 5 and 6. Above analysis can be outlined in following points.

- FP-Growth algorithm outperforms on low support count.

- Apriori performs better on the high support count.

- FP-Growth takes much processing time, if large transaction set is given, due to its requirement of large storage memory in order to store a tree structure.

TABLE I.   EXECUTION TIME OF APRIORI AND FP-GROWTH ON DIFFERENT SUPPORT COUNTS (FOR T10I4D100K DATASET)

Support Count	Time (MS)	
	FP-Growth	Apriori
20	26900	46907
30	3737	7860
40	1206	2301
50	1126	929
60	1009	604
70	973	613
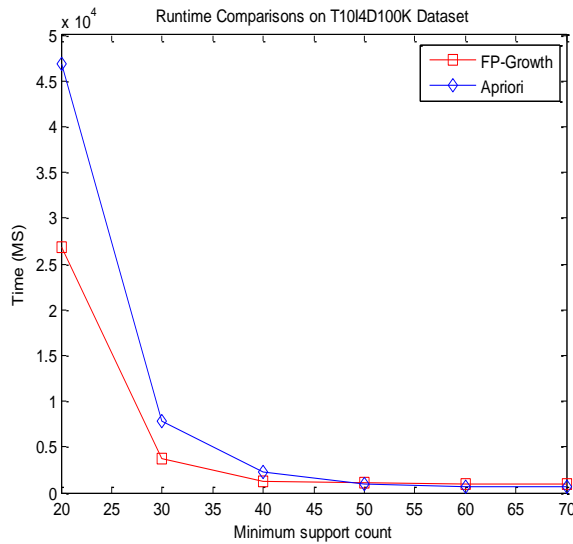
Fig. 5.   Runtime comparative analyses on T10I4D100K Dataset

TABLE II.   EXECUTION TIME OF APRIORI AND FP-GROWTH FOR PIMA D38.N768.C2 DATASET

Support Count	Time(MS)	
	FP-Growth	Apriori
10	51	94
20	42	78
30	42	72
40	35	55
50	30	45
60	25	32
70	19	25
80	16	16
90	8	7
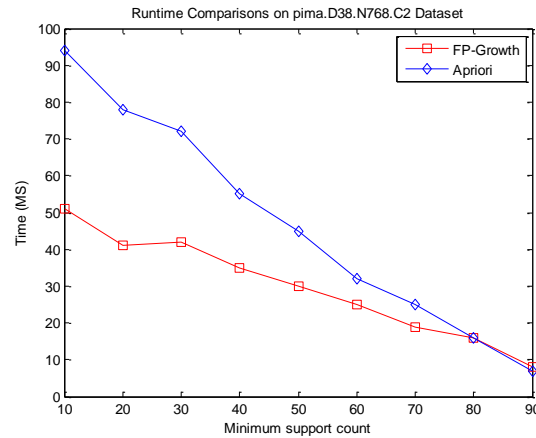
Fig. 6.   Runtime comparative analyses on pima.D38.N768.C2 Dataset

VI. PROPOSED APPROACH

To overcome the limitations of the previous AIRM approach we adopt the concept of fuzzy clustering instead of classification used in previous (AIRM) approach which accepts frequent patters with degree of matching 1 (100%), whereas the newly proposed algorithm accepts all the frequent patterns either it has degree of matching 0, 1 or between of 0 & 1 because of its fuzzy nature. Under this process FCM accepts the normalized data set which contained normalized degrees of matching and produces clusters to demonstrate the nature of frequent patterns. In order to implement a fuzzy inference mechanism as shown in figure.7, AFIRM follows the following steps.

A. Data selection and Preprocessing

This is the very first step of AFIRM algorithm where data selection has performed for preprocessing, this step first load and scan the main dataset as well as fact dataset, then preprocess the datasets by mapping items/ symptoms by its corresponding index values. For this it previously read all items/symptoms and create an index table by assigning a unique index value to each item/symptom.

B. Association rule mining

This step performs the association rule mining in order to find frequent patterns, hence we use the FP - growth algorithm because it reads database once, and find frequent patterns. This process generates all the frequent patterns or frequent symptoms as per given thresholds.

C. Fact matching and matrix generation

This is the most significant step of this algorithm where all the frequent symptoms match with fact dataset and find its degree of similarity (where factual data set contain symptoms with related disease), afterward all these information further stores in another database where frequent symptoms, disease and degree of similarity camps to gather in the form of a matrix.

D. Matrix normalization

Once the matrix of DOM is completely formed, a normalization step is conceived to normalize the degree of similarity for applying fuzzy, c-means algorithm, here all the degrees must be between 0 and 1. If the degree is 0 means no similarity is found and if degree is 1 means frequent symptom fully similar to disease symptoms and if the degree is between 0.01 to 0.99 mean partial similarity has found.

Index Table

```
0,ulcers_of_the_mouth
1,loss_of_appetite
2,vaginal_candidiasis
3,nasal_secretions
4,fungal_nails
5,vomiting
6,diarrhea
7,vaginal_pain
8,hunger
9,chills
10,frequent_urination
11,watery_eye
12,cough
13,shortness_of_breath
14,nipple_discharge
15,abdominal_pain
16,rash
17,sore_throat
18,fungal_infection_of_the_mouth_and_
esophagus
19,headache
20,joint_pain
21,breast_pain_nipple_pain
```

Clinical Dataset
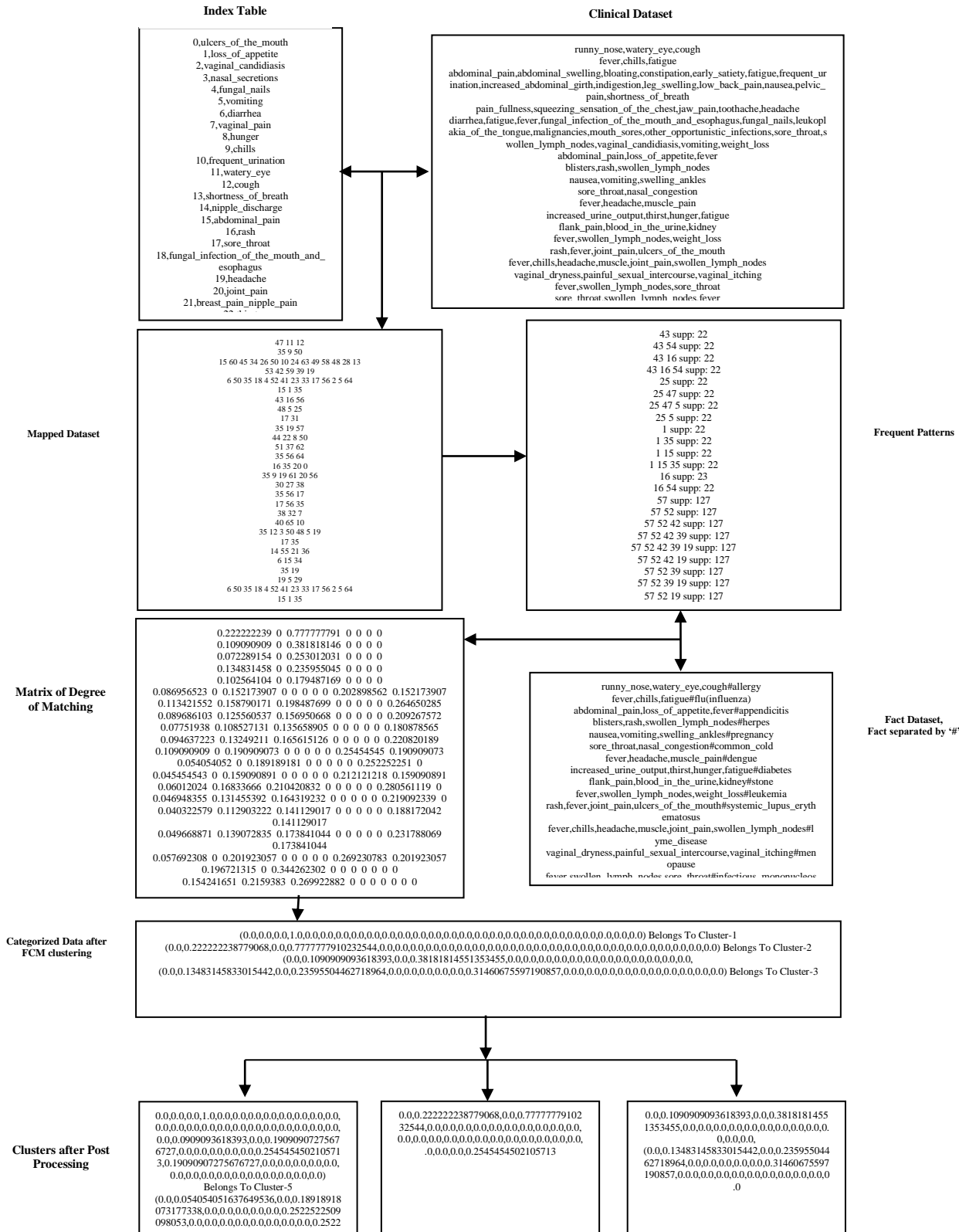
```
runny_nose,watery_eye,cough
fever,chills,fatigue
abdominal_pain,abdominal_swelling,bloating,constipation,early_satiety,fatigue,frequent_ur
ination,increased_abdominal_girth,indigestion,leg_swelling,low_back_pain,nausea,pelvic_
pain,shortness_of_breath
pain_fullness,squeezing_sensation_of_the_chest,jaw_pain,toothache,headache
diarrhea,fatigue,fever,fungal_infection_of_the_mouth_and_esophagus,fungal_nails,leukopl
akia_of_the_tongue,malignancies,mouth_sores,other_opportunistic_infections,sore_throat,s
wollen_lymph_nodes,vaginal_candidiasis,vomiting,weight_loss
abdominal_pain,loss_of_appetite,fever
blisters,rash,swollen_lymph_nodes
nausea,vomiting,swelling_ankles
sore_throat,nasal_congestion
fever,headache,muscle_pain
increased_urine_output,thirst,hunger,fatigue
flank_pain,blood_in_the_urine,kidney
fever,swollen_lymph_nodes,weight_loss
rash,fever,joint_pain,ulcers_of_the_mouth
fever,chills,headache,muscle,joint_pain,swollen_lymph_nodes
vaginal_dryness,painful_sexual_intercourse,vaginal_itching
fever,swollen_lymph_nodes,sore_throat
sore_throat,swollen_lymph_nodes,fever
```

Mapped Dataset

```
47 11 12
35 9 50
15 60 45 34 26 50 10 24 63 49 58 48 28 13
53 42 59 39 19
6 50 35 18 4 52 41 23 33 17 56 2 5 64
15 1 35
43 16 56
48 5 25
17 31
35 19 57
44 22 8 50
51 37 62
35 56 64
16 35 20 0
35 9 19 61 20 56
30 27 38
35 56 17
17 56 35
38 32 7
40 65 10
35 12 3 50 48 5 19
17 35
14 55 21 36
6 15 34
35 19
19 5 29
6 50 35 18 4 52 41 23 33 17 56 2 5 64
15 1 35
```

Frequent Patterns

```
43 supp: 22
43 54 supp: 22
43 16 supp: 22
43 16 54 supp: 22
25 supp: 22
25 47 supp: 22
25 47 5 supp: 22
25 5 supp: 22
1 supp: 22
1 35 supp: 22
1 15 supp: 22
1 15 35 supp: 22
16 supp: 23
16 54 supp: 22
57 supp: 127
57 52 supp: 127
57 52 42 supp: 127
57 52 42 39 supp: 127
57 52 42 39 19 supp: 127
57 52 42 19 supp: 127
57 52 39 supp: 127
57 52 39 19 supp: 127
57 52 19 supp: 127
```

Matrix of Degree of Matching

```
0.222222239 0 0.777777791 0 0 0 0
0.109090909 0 0.381818146 0 0 0 0
0.072289154 0 0.253012031 0 0 0 0
0.134831458 0 0.235955045 0 0 0 0
0.102564104 0 0.179487169 0 0 0 0
0.086956523 0 0.152173907 0 0 0 0 0.202898562 0.152173907
0.113421552 0.158790171 0.198487699 0 0 0 0 0.264650285
0.089686103 0.125560537 0.156950668 0 0 0 0 0.209267572
0.07751938 0.108527131 0.135658905 0 0 0 0 0.180878565
0.094637223 0.13249211 0.165615126 0 0 0 0 0.220820189
0.109090909 0 0.190909073 0 0 0 0 0.25454545 0.190909073
0.054054052 0 0.189189181 0 0 0 0 0 0.252252251 0
0.045454543 0 0.159090891 0 0 0 0 0.212121218 0.159090891
0.06012024 0.16833666 0.210420832 0 0 0 0 0.280561119 0
0.046948355 0.131455392 0.164319232 0 0 0 0 0.219092339 0
0.040322579 0.112903222 0.141129017 0 0 0 0 0.188172042
0.141129017
0.049668871 0.139072835 0.173841044 0 0 0 0 0.231788069
0.173841044
0.057692308 0 0.201923057 0 0 0 0 0.269230783 0.201923057
0.196721315 0 0.344262302 0 0 0 0 0 0
0.154241651 0.2159383 0.269922882 0 0 0 0 0 0
```

Fact Dataset, Fact separated by '#'

```
runny_nose,watery_eye,cough#allergy
fever,chills,fatigue#flu(influenza)
abdominal_pain,loss_of_appetite,fever#appendicitis
blisters,rash,swollen_lymph_nodes#herpes
nausea,vomiting,swelling_ankles#pregnancy
sore_throat,nasal_congestion#common_cold
fever,headache,muscle_pain#dengue
increased_urine_output,thirst,hunger,fatigue#diabetes
flank_pain,blood_in_the_urine,kidney#stone
fever,swollen_lymph_nodes,weight_loss#leukemia
rash,fever,joint_pain,ulcers_of_the_mouth#systemic_lupus_eryth
ematosus
fever,chills,headache,muscle,joint_pain,swollen_lymph_nodes#l
yme_disease
vaginal_dryness,painful_sexual_intercourse,vaginal_itching#men
opause
fever,swollen_lymph_nodes,sore_throat#infectious_mononucleos
```

Categorized Data after FCM clustering

```
(0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) Belongs To Cluster-1
(0,0,0.222222238779068,0,0,0.7777777910232544,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) Belongs To Cluster-2
(0,0,0.1090909093618393,0,0,0.38181814551353455,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
(0,0,0.13483145833015442,0,0,0.23595504462718964,0,0,0,0,0,0,0,0,0,0,0,0,0.31460675597190857,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) Belongs To Cluster-3
```

Clusters after Post Processing

```
0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0.0909093618393,0,0,0.19090907275676
727,0,0,0,0,0,0,0,0,0,0,0.254545450210571
3,0.19090907275676727,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
Belongs To Cluster-5
(0,0,0.05405405163764953 6,0,0,0.18918918
073177338,0,0,0,0,0,0,0,0,0,0,0.2522522509
098053,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.2522
```

```
0,0,0.222222238779068,0,0,0.77777779102
32544,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
,0,0,0,0,0.2545454502105713
```

```
0,0,0.1090909093618393,0,0,0.3818181455
1353455,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,
(0,0,0.13483145833015442,0,0,0.235955044
62718964,0,0,0,0,0,0,0,0,0,0,0.31460675597
190857,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
.0
```

Fig. 7. Proposed Framework for AFIRM

E. FCM Clustering and Inference evaluation

FCM is the fuzzy based clustering algorithm so that it can better categorize the above generated degrees of similarity, hence we apply the Fuzzy c-means algorithm on normalized matrix in order to categorize the frequent symptoms. When we plot these clusters, we seem that visualization, demonstrate that which diseases is highly effected in a session.

VII. ALGORITHM: ASSOCIATION FUZZY INFERENCE RULE MINER (ABBREVIATED AS AFIRM)

TABLE III.   ABBREVIATIONS USED:

Abbreviation	Meaning
DS	Dataset
FP	Frequent patterns
ARM	Association rule mining
Infr	Inference
Min_sup	Minimum support
FIS	Fuzzy inference system
FCM	Fuzzy C-Means
DMMTX	Degree of Matching Matrix
NDMMTX	Normalize Degree of Matching Matrix

1. Begin:

2. Load (Dataset), Min_Sup_Threshold, Fact_DS

   //Data Preprocessing

3. Index=Gen_Index(Dataset)

4. Mapped_DS=Mapping(Dataset, Index)

   //Association Rule Mining

5. FP=ARM(Mapped_DS, Min_Sup_Threshold)

   // FP-Growth By Han et.al [20]

6. DMMTX=Match(FP, Fact_DS)

   // Fact Matching & DOM Matrix Implementation

7. NDMMTX=Normalization(DMMTX)

   //Normalization of DOM matrix

8. Infr_Clusters=FCM(FP, Fact_DS)

9. PostProcessing(Infr_Clusters)

10. Result_Analysis

11. End

Procedure: FCM (NDMMTX)

1. Begin

2. Initialize matrix U for degree of membership U=[uij]

3. At each step calculate centers vector C (k) & U (k).

4. Calculate the d-dimension center of the cluster.

5. Update U(k), U(k+1)

6. Repeat step 2 till MAXij< termination decisive factor

7. End.

A. Case Study

After preprocessing, original dataset will map in following form.

```
 1  46 11 12
 2  35 9 49
 3  15 58 45 34 26 49 10 24 61 48 56 47 28 13
 4  52 42 57 39 19
 5  6 49 35 18 4 51 41 23 33 17 54 2 5 62
 6  15 1 35
 7  43 16 54
 8  47 5 25
 9  17 31
10  35 19 55
11  44 22 8 49
12  50 37 60
13  35 54 62
14  16 35 20 0
15  35 9 19 59 20 54
16  30 27 38
17  35 54 17
18  17 54 35
19  38 32 7
20  40 63 10
21  35 12 3 49 47 5 19
22  17 35
23  14 53 21 36
24  6 15 34
25  35 19
26  19 5 29
27  6 49 35 18 4 51 41 23 33 17 54 2 5 62
28  15 1 35
29  43 16 54
30  47 5 25
31  17 31
32  6 49 35 18 4 51 41 23 33 17 54 2 5 62
33  15 1 35
34  43 16 54
35  47 5 25
36  17 31
37  6 49 35 18 4 51 41 23 33 17 54 2 5 62
38  15 1 35
39  43 16 54
40  47 5 25
```

Fig. 8.   Mapped dataset

After applying ARM we got 511 frequent symptoms on 40% of minimum support.

```
ARM

47 supp: 140
47 5 supp: 140
23 supp: 140
0 supp: 140
0 42 supp: 140
0 42 28 supp: 140
0 28 supp: 140
26 supp: 151
10 supp: 151
45 supp: 163
45 28 supp: 140
45 28 9 supp: 140
45 9 supp: 140
42 supp: 165
42 5 supp: 141
42 5 28 supp: 141
42 28 supp: 143
1 supp: 259
5 supp: 282
```
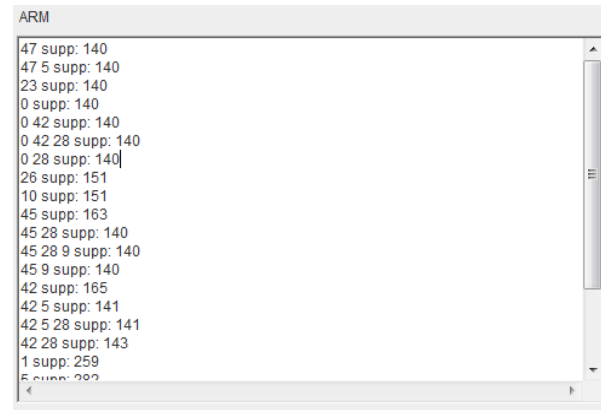
Fig. 9.   Frequent pattern discovery under ARM

Figure.11 shows the resulting clusters, which categorize the frequent symptoms on the basis of degree of matching (DOM) in order to demonstrate and categorize the possibility of disease infection. Earlier than, DOM normalization has applied to transform the DOM in the required form, as shown in Figure-10.

After clustering a post processing phase is needed to cluster the data and validate the result, it depends on data properties, where data points are plotted with different colors depending on the cluster assigned. In this step we apply the post processing on clusters, to obtain the exact information about clusters.
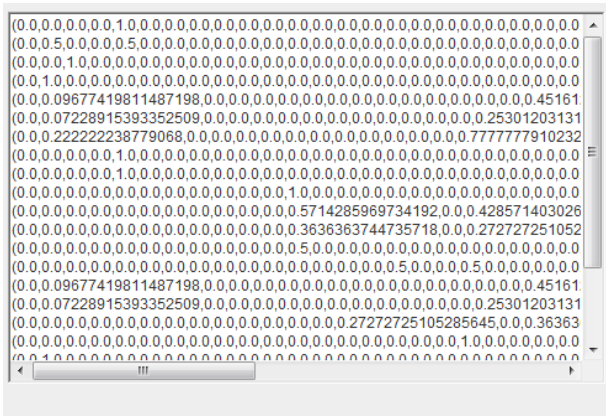
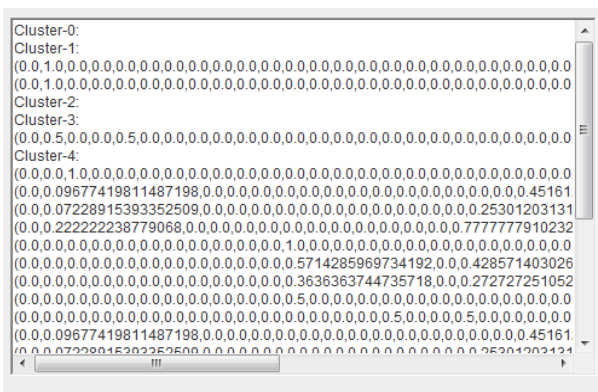Fig. 10. Normalized (Pre-processed) matrix for FCM

Fig. 11. Clusters generated by FCM algorithm

VIII. SIMULATION AND RESULT

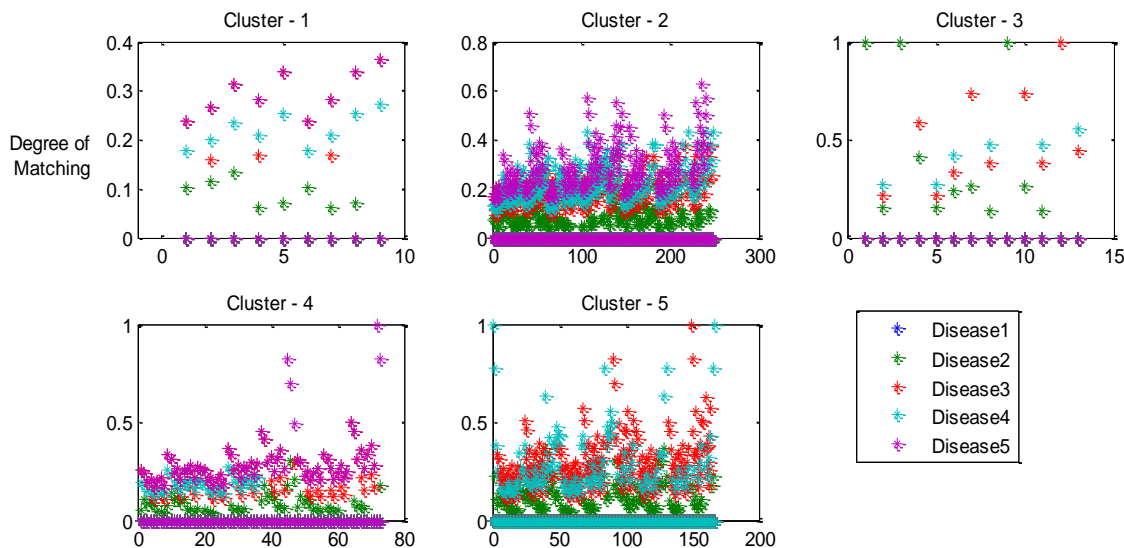Figure 12 represents a graphical representation of clusters generated by fuzzy, c-means algorithm, this scatters view,

representing the amount of possible disease in a particular cluster with the respective degree of matching. In this figure cluster 2 and 4 contain the high number of objects as diseases and showing that the disease5 has a high possibility of infection under this study, we also infer that the cluster 2, 4 and 5 are represented which disease would be highly effective.

IX. CONCLUSION

In this paper, we have proposed a fuzzy, c-means based inference system for association rule mining, in this study, we propose a novel algorithm AFIRM (Association fuzz inference rule mining) which is an extended version of previously proposed algorithm AIRM.

It consists of two phases. First phase scan the given dataset with corresponding fact dataset and perform preprocess to meet required format for rule mining then we apply FP-growth on pre-processed dataset in order to find frequent patterns. Second, we match these frequent patterns with fact dataset and create a matrix of degree of matching or similar and then, we normalize this matrix to apply FCM procedure. Therefore Fuzzy c-means clustering algorithm categorizes this data in different clusters on the basis of pre assigned degree of matching. Experimental results show that FIRM (Fuzzy inference rule miner) out performs comparisons to previous approach AIRM in order to discover background knowledge more over use full inference.

In future AFIRM can also use the concept of Markov predictor to know future possibilities. Secondly AIRM and AFIRM both algorithms first discover the association rules using FP-growth algorithm, to optimize, an expert mechanism might be explored to find most suitable and efficient algorithm instead of FP-growth algorithm on the basis of the type of dataset to efficiently discover association rules.

Fig. 12. Clusters generated after FCM, representing amounts of possible diseases in resulting clusters

REFERENCES

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." In Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, (1994): pp. 487-499.

[2] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." In ACM SIGMOD Record, vol. 22(2), (1993): pp.207-216.

[3] Bezdek, James C., "Pattern recognition with fuzzy objective function algorithms", Kluwer Academic Publishers, 1981.

[4] Chaturvedi, Kapil, Dr. Ravindra Patel and Dr. D.K. Swami, "An Inference Mechanism Framework for Association Rule Mining" International Journal of Advanced Research in Artificial Intelligence(IJARAI), vol. 3(9), (2014): pp.1-8.

[5] Chazard, Emmanuel, Gregoire Ficheur, Stephanie Bernonville, Michel Luyckx, and Regis Beuscart. "Data mining to generate adverse drug events detection rules." Information Technology in Biomedicine, IEEE Transactions, vol. 15(6), (2011): pp.823-830.

[6] Chow, Richard, Philippe Golle, and Jessica Staddon. "Detecting privacy leaks using corpus-based association rules." In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, (2008): pp. 893-901.

[7] Cominetti, Ornella, Anastasios Matzavinos, Sandhya Samarasinghe, Don Kulasiri, Sijia Liu, Philip Maini, and Radek Erban. "DifFUZZY: a fuzzy clustering algorithm for complex datasets." International Journal of Computational Intelligence in Bioinformatics and Systems Biology vol. 1(4), (2010): pp. 402-417.

[8] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." Journal of cybernetics, vol. 3, (1973): 32-57.

[9] Fagin, Ronald, Joseph Y. Halpern, and Moshe Y. Vardi. "What is an inference rule?." The Journal of symbolic logic, vol. 57(03), (1992): pp.1018-1045.

[10] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine, vol. 17(3), (1996): pp.1-34.

[11] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." ACM SIGMOD Record, vol. 29(2), (2000).

[12] http://fimi.ua.ac.be/data/T10I4D100K.dat (Frequent Itemset Mining Dataset Repository).

[13] http://www.cdc.gov/nchs/icd/icd9.html

[14] http://www.cs831/notes/itemsets/datasets.php (the lucs-kdd software library)

[15] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco, USA, ISBN 1558604898, 2001.

[16] Ji, Yanqing, Hao Ying, Peter Dews, Ayman Mansour, John Tran, Richard E. Miller, and R. Michael Massanari. "A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance."Information Technology in Biomedicine, IEEE Transactions, vol. 15(3), (2011): pp.428-437.

[17] Kaufman, Leonard, and Peter J. Rousseeuw. "Partitioning around medoids (program pam)." Finding groups in data: an introduction to cluster analysis (1990): pp. 68-125.

[18] Kaushik, Saroj, "Artificial Intelligence", Cengage Learning, India, ISBN.9788131510995, 2011.

[19] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, (1967): pp. 281-297.

[20] Nguyen, Thanh, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi, "Medical diagnosis by fuzzy standard additive model with wavelets." Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on. IEEE, (2014): pp.1937-1944.

[21] Samawi, Venus , Akram Mustafa, and Abeer Ahmad. , "Arabic Expert System Shell", IAJIT First Online Publication, (2011): vol. 10(1).

[22] Tao, Xuehong, Yuan Miao, Yanchun Zhang, and Zhiqi Shen. "Collaborative medical diagnosis through Fuzzy Petri Net based agent argumentation." InFuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on, (2014): pp.1197-1204.

[23] Wang, Xiaozhe, Ajith Abraham, and Kate A. Smith. "Intelligent web traffic mining and analysis." Journal of Network and Computer Applications, Vol. 28(2), (2005): pp.147-165.

[24] Yang, Jian-Bo, Jun Liu, Jin Wang, How-Sing Sii, and Hong-Wei Wang. "Belief rule-base inference methodology using the evidential reasoning approach-RIMER." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions, vol. 36(2), (2006): pp.266-285.

[25] Yu, Zhiwen, Zhiyong Yu, Xingshe Zhou, Christian Becker, and Yuichi Nakamura. "Tree-based mining for discovering patterns of human interaction in meetings." Knowledge and Data Engineering, IEEE Transactions, vol. 24(4), (2012): pp.759-768.

AUTHOR PROFILE

Mr. Kapil Chaturvedi has post graduated in Master of Computer Application (MCA) from from Samrat Ashok Technical Institute (Degree), Vidisha, India. Presently he is pursuing Ph.D. from Rajiv Gandhi Technological University, (Bhopal) India. He has more than 6 year teaching experience in post-graduation, He has published more than 6 research papers in international and conference proceedings He is working as Assistant Professor in Department of Computer Applications, UIT, Rajiv Gandhi Technical University, Bhopal, India, his area of interest is Data Mining & Digital Image Processing.

Dr. Ravindra Patel, Associated Professor and Head, Department of Computer Applications at Rajiv Gandhi Technological University, (Bhopal) India. He is Ph.D. in Computer Science. He possesses more than 10 years of teaching experience in post-graduation. He has published more than 30 research papers in international and national journals, and conference proceedings. He is member of ISTE and International Association of Computer Science and Information Technology (IACSIT).

Dr. D. K. Swami is presently Group Director, VNS Group of Institutions, Bhopal. Before joining VNS in 2008, he has served Samrat Ashok Technological Institute (SATI), Vidisha for 20 years. He obtained his Ph. D. in Computer Science and Engineering in 2007 from Rajiv Gandhi Technological University, Bhopal. He completed M. Tech in Computer Applications from IIT, Delhi in 1994 and he completed M.Sc. in Applied Mathematics in 1988. During 26 years of teaching, he has taught various courses in computer science to the students of BE, M. Tech. and MCA students. He has more than 30 research publications I journals and in conference proceedings of repute. He has edited one book and authored a book on Basic Computer Engineering. He is a member for life ISTE and senior life member of CSI. Areas of his special interest include Data Mining, DBMS, Object Oriented System and software Engineering.