International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue III, Mar 2019- Available at www.ijraset.com

Hybrid System to find Association using Genetic Algorithm and Fuzzy Logic

Archana Gupta[1], Parakh Shah[2], Deep Mehta[3]
[1]*Assistant Professor, K.J.S.C.E., Department of Computer Engineering, Vidyavihar, Mumbai*
[2, 3]*Student, K.J.S.C.E., Department of Computer Engineering, Vidyavihar, Mumbai*

Abstract: ***Now a day, so many technologies are evolving around to find solution of the problem. Association rule mining is a well-known technology used to find the association among the different attributes in any field. ARM cannot work directly on numerical values, so concept of fuzzy logic will applied to assign Linguistic variable. And Genetic algorithm is an evolutionary algorithm to get the optimized result. GA is based on the theory of survival. Fittest chromosome has high probability of surviving. In this paper, author used the concept of genetic algorithm in association rule mining to get the frequent itemset. And before applying genetic algorithm, data has to be fuzzify to support ARM algorithm. With the help of this implementation rare and popular, both cases will be considered to derive the associations among the parameters which can be used further in any decision making strategy.***

I. INTRODUCTION

Data Mining in itself is a package with many algorithms well suited to solve complex problems. Data mining algorithms can be implemented for prediction, classification, pattern detection, finding association, series analysis, etc on different types of data such as temporal, spatial, geospatial, day to day data and many more. Now a day, data is exploded in huge amount in every application and data mining algorithms has broad area of application with variety of data. Though in Data Mining, Association rule extraction is the most widely based exploration technology and mainly used to find hidden relationships between data in order to generate classification clusters, wherein data items are combined based on their various granularity levels. The most popular example of association rule extraction is market basket analysis. Data Mining techniques are used to make decisions in business strategy, given that traditional methods of Big data analysis have become inefficient and show poor performance.

Association rule mining [1] can be used in many areas to predict the association among the influencing factor of that application. In medical science, we can predict the association of different parameters with its values for the chance of having any disease. This will help to control the factors so that the disease will not occur.

The remaining of this paper is organized as: section 2 gives the introduction to the basic concept of association rule mining, Genetic algorithm and Fuzzy set theory, section 3 gives the related work in the same field, section 4 gives the hybrid system used in association rule mining using genetic algorithm and fuzzy set theory for the chance of having heart disease, section 5 gives the experimentation results of the implemented system, finally, the system will be concluded along with future work in section 6.

II. BASIC CONCEPTS

A. Association Rule Mining (ARM)

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Initially Market basket analysis [1] is the major application of ARM. Based on the concept of strong rules, association rules can be formulated for discovering regularities between products in transactional data recorded by systems in supermarkets. For example, the rule {onion, potato}->{burger bun} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy burger buns. Such information can be used as the basis for decisions making about marketing activities such as, e.g., promotional pricing or product placements. Now a day, ARM can be used in many application areas including Web usage mining, intrusion detection, continuous production, bioinformatics and many more.

Features used in ARM are:

Support:

Support is an indication of how frequently the itemset appears in the dataset.

The support of X with respect to T is defined as the proportion of transactions t in the dataset which contains the itemset X.

Supp(X) = Total number of tractions having X / Total number of transactions.

Confidence:

Confidence is an indication of how often the association rule has been found to be true with respect to given transactional database.

The confidence value of an association rule, X ⇒ Y, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.

Confidence is defined as:

Conf (X ⇒ Y)= Supp(X ∪ Y)/Supp(X)

B. Genetic Algorithm

In computer science, a genetic algorithm (GA) [6] is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA). Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection. In a genetic algorithm, a population of candidate solutions (called individuals) to an optimization problem is evolved toward better solutions. Solutions are normally represented in binary as strings of 0s and 1s

A typical genetic algorithm requires:

a) a genetic representation of the solution domain,

b) a fitness function to evaluate the solution domain.

C. Fuzzy System

Fuzzy logic is a form of many-valued logic in which the truth values of variables may be any real number between 0 and 1 inclusive. It is employed to handle the concept of partial truth [9], where the truth value may range between completely true and completely false. It is based on the observation that every information cannot be very precise and numerical, fuzzy models are mathematical means of representing vagueness and imprecise information, hence the term fuzzy.

In the Process to generating a fuzzy system, we need to fuzzify all input values using fuzzy membership functions, apply Fuzzy rule base to infer output. It may require to De-fuzzify the fuzzy output value to get "crisp" output value.

So here we need methods to do Fuzzification, Defuzzification and to generate fuzzy rules.

Under Fuzzification, Fuzzify the numerical data to give meaning to the linguistic variable by using some fuzzification method such as triangle or trapezoid-shaped curves,etc. One example is as below.

In defuzzification get the crisp output by analysing the fuzzy rules using some defuzzification method[2].

III. RELATED WORK

Rajdeep Kaur Aulakh et al. [13] In this paper initially Apriori algorithm is applied, in order to generate frequent item-sets and then frequent item-sets are used to generate association rules. After getting association rules from Apriori algorithm Genetic Algorithm (GA) is applied to obtain reduced number of association rules. A new fitness function is proposed for the application of Genetic algorithm. It is observed that this algorithm greatly reduces the problem of generation of huge association rules using Apriori algorithm. The implementation of the proposed algorithm is easier than other popular algorithm for association rule mining. The proposed algorithm performs much better when compared to Apriori algorithm and other previous technique used to optimize association rule mining. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database.

Ayush Kumar Agrawal et al. [14] In this paper, study of various research to find how to mine the large item set is done. Along with that study is done to find the way to reduce it and to know how to design fuzzy association rule for better mining. There are various technique available that are used to mine dataset of large item set but fuzzy mining association rule is better than others. The advantages and the limitations of fuzzy mining association rule is also discussed in this paper. In this paper, the main aim is to study the features and importance of fuzzy mining association rule technique and their advantages and disadvantages.

Basheer Mohamad Al-Maqaleh et al. [15] , proposed a multi-objective genetic algorithm approach in this paper for the discovery of interesting association rules with multiple criteria i.e. support, confidence and simplicity (comprehensibility). In this paper a global search can be achieved with Genetic Algorithm (GA), and system automation is developed. The large number of rules generated by the Apriori algorithm makes manual inspection of the rules very difficult. It is hence impossible for an expert of the field being mined to sustain these rules. Thus GA is used here which considerably outperformed the Apriori algorithm in datasets, with respect to the number of discovered rules.

Mimanshu Gupta et al. [16] study on how to reduce large data sets to smaller data set by applying fuzzy logic association rules and also to forecast from large data set. They integrates data mining with Fuzzy Logic. Implementation of the fuzzy logic helps in reduction of data. Here they presented the literature review in the field of fuzzy mining association rules using different technology and found that still we can predict the result from large data set.

IV. PROPOSED HYBRID SYSTEM

In this section, Hybrid system is generated in the field of medical science to predict that what are different factors that have strong associations to cause heart disease. If strong association among the factors are known and if a pateint is already suffered from some symptoms then precaution can be taken to control that factor that are likely to occur. So this work will help in reducing the chance of having heart disease. Association rule mining is a very good approach to find the association among the various factors. Here, authors are trying to generate an hybrid system to combine Fuzzy logic and Genetic algorithm with Association rule mining.
The block diagram of the same hybrid system is give in Fig 1

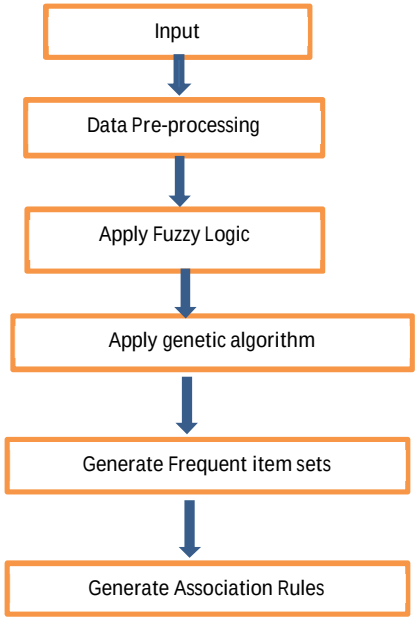

Fig 1. Hybrid system to generate Association rules

A. Original Dataset

The dataset consists of 300 individuals data. There are 15 columns in the dataset, however the first column name is not a good parameter as far as machine learning is considered so, there are effectively 14 columns named: Age, Sex, Chest-pain type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar , Resting ECG , Max heart rate achieved , Exercise induced angina, ST depression induced by exercise , Peak exercise ST segment, Number of major vessels, Thal, Diagnosis of heart disease.
The above mentioned is the original data set that are available for the heart disease prediction at various sites. In this paper, a system is designed to find the correlation among the common factors that are understandable by the general community. So the original dataset is reduced here and only 6 commonly known attributes are used to find the association and present the work. The 6 used attributes are – Age, Cholesterol, Blood pressure, Blood Sugar Level, Maximum Heart rate, ST depression induced by Exercise.

B. Pre-processing of Data

It is the second phase of the proposed hybrid system. Data pre-processing is required to clean the data and to check that data has assigned meaning or not. Sometimes it is required to fill the missing values as well in the dataset. Here missing values are handles using logistic regression.
Logistic regression:
Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

C. Fuzzy Logic

In third phase of Apply Fuzzy Logic, Fuzzification of the different factors will be based on membership functions and Rule base will be generated for the output "Chance of having Heart Disease".

Out of the given 13 factors of the dataset of heart disease, 5 factors are having continuous values thus it is required to convert these continuous values into linguistic variables. The membership functions of the different linguistic variable used for these factors are given in subsequent section.

Based on the study of the available dataset, the following membership function has been decided for the above mentioned five parameters out of 13 given parameters:

1) Age: Age is the most important risk factor. This input variable has divided to 3 fuzzy sets described as below

Input Field	Fuzzy set	Range	Membership function
Age	Low	30-50	Trapezoidal
	Medium	45-65	Triangular
	High	63onwards	Trapezoidal

2) Blood Pressure: In this field, we use systolic blood pressure. This input variable has divided to 4 fuzzy sets described in [12] as below :

Input Field	Fuzzy set	Range	Membership function
Systolic blood pressure	Low	<134	Trapezoidal
	Medium	127-153	Triangular
	High	142-172	Triangular
	Very high	>154	Trapezoidal

3) Cholesterol: For this input field, we use the value of low density lipoprotein (LDL) cholesterol. Cholesterol field has 4 fuzzy [12] as below:

Input Field	Fuzzy set	Range	Membership function
Cholesterol	Low	<197	Trapezoidal
	Medium	188-250	Triangular
	High	217-307	Triangular
	Very high	>281	Trapezoidal

4) Maximum Heart Rate: The value of this field is maximum heart rate of man in 24 hours. By increasing of age in man, maximum of heart rate in 24 hours decreases. In this field, we have 3 linguistic variables [12] fuzzy sets as given below:

Input Field	Fuzzy set	Range	Membership function
Maximum heart rate	Low	<141	Trapezoidal
	Medium	111	Triangular
	High	194	Triangular
	Very high	>152	Trapezoidal

5) ST Depression: In this field, we have 3 fuzzy sets (Normal, ST_T abnormal & Hypertrophy) [12] as given in the below table:

Input Field	Fuzzy set	Range	Membership function
Resting Electrocardiography (ECG)	Normal	[-0.5,0.4]	Trapezoidal
	ST-T abnormal	[0.45,1.8]	Triangular
	Hypertrophy	[1.4,2.5]	Trapezoidal

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue III, Mar 2019- Available at www.ijraset.com

D. Apply Genetic Algorithm

In the fourth phase of Apply Genetic algorithm, Initial Population of chromosomes is generated by using inference rules generated in second phase. Then after applying selection, crossover and mutation, new population will get generate. This will go for many iteration until stopping criteria met.

After deciding for the linguistic variables for all the factors, the next step is to design the rule base of the fuzzy system.

So, here some cases of the rule base [2] to have the chance of having the heart disease.

Some sample fuzzy rules are:

Age	Resting Blood pressure	Serum Cholesterol	Fasting Blood Sugar	Maximum heart rate	ST depression
L	L	L	1	L	N
M	M	M	0	M	SA
H	H	H	1	H	H
L	VH	VH	0	VH	N
M	L	L	1	L	SA
H	M	M	0	M	H
L	H	H	1	H	N
M	VH	VH	0	VH	SA
H	L	L	1	L	H
L	M	M	0	M	N

These rule bases will give the association among the values of the different factors used to predict the chance of having heart disease.

After getting the association rules, genetic algorithm is used to get the optimized association rules. To implement genetic algorithm there are two important requirements:

1) How to represent the chromosomes
2) How to define the fitness function

a) Chromosome Representation: Binary encoding is used to represent chromosomes. Since there are six factors and below is the number of bits required to represent a factor in chromosome based on the linguistic variables.

Since Age has 3 different values, Sex has 2 different values, Chest pain type has 4 different values, Resting Blood pressure has 4 different values, Serum Cholesterol has 4 different values, Fasting Blood Sugar has 2 different values, Resting ECG has 3 different values , Maximum heart rate has 4 different values, Exercise induced angina has 2 different values , ST depression has 3 different values , peak exercise ST segment has 3 different values , Number of major vessels has 4 different values , thal has 3 different values (1-3,2-6,3-7)

A	RBP	S C	FBS	MHR	STD
L	L	L	1	L	N
	00 00	00	1	00	00

b) Fitness Function: Fitness function used is the actual support of the chromosome in the given input sample database. Blood sugar level is not considered while evaluating the fitness of the chromosome.

Fitness of chromosome C1 = $\sum^{i=1 \text{ to } n}$(min (μ(k) for the said linguistic variable for all the factors from k = 1 to 6 in tuple(i)))

Where n is the total number of tuples in sample database.

Genetic Algorithm is implemented in the following way:

i) Create initial population, from the above mentioned fuzzy rules which is given in above table.
ii) Randomly select parents for crossover
1. For selection, calculate fitness function for every individual in population.
2. Generate a random number in between min and max of fitness value.
3. Two chromosomes that are closest to the generated random number will get select.

iii) Then perform crossover by randomly selected crossover points out of points given above. After crossover check for the validity of all six component of the children chromosome for example 11 is not valid for maximum heart rate. If it happens then discard the crossover and select next closest chromosome (as given in 2.c), change only one parent and then repeat 3.

iv) Then again generate random number for all 6 components of the chromosome. If random number is greater than 0.5 then perform mutation.

1. For single length component mutation is just flip the bit(0 to 1 and 1 to 0)
2. But for two bits component , randomly select first bit or second bit (by randomly generate either 0 or 1), and then flip it. And after flipping check the validity as well for example 11 is not valid for maximum heart rate.

v) Perform crossover with 5 pairs of the parents, then 10 children will get generate.

vi) Now we will have two sets of the chromosomes, first set comprises of the chromosomes defined in step 1 and the second set consist of all the chromosomes we will get after applying crossover and mutation. Calculate the fitness function value of all 20 chromosomes. Select randomly 10 chromosomes based on the fitness value to place in the next population / generation.

vii) Then repeat step 2 – 6 for n number of times, n will be decided by the difference in the maximum fitness value in previous and present generation. If difference is less than 0.001 then stop.

After giving the initial population, then crossover and mutation operation will be applied to get the next population. And then the best chromosomes based on their fitness value are more likely to be present in next population. The entire iteration will get repeat until stopping criteria met. Finally, a set of frequent itemsets can generate from final population, which can be further used to generate association rules.

E. In the fifth phase list of frequent itemsets will get generate with the help of fit chromosomes in the population.

F. In the sixth phase Association rules will get generate from the frequent itemset which is not shown in this paper it is kept in future work and this will be done again by using genetic algorithm as it is known that number of association rules increase exponentially with the size of frequent itemset. So if Association rules will get generate using traditional approach of Apriori Algorithm, it will cost a lot. Thus genetic algorithm is proposed to optimize the work of association rule generation.

V. EXPERIMENTATION RESULTS

To evaluate the performance of the proposed system Extensive simulation experiments have been performed. The goal is to optimize the processes of finding frequent itemsets. In this research work, the performance of proposed system is compared with that of Apriori algorithm.

A. Results on Heart Disease Dataset

We applied our proposed algorithm on Heart Disease dataset. We obtained 10 frequent itemsets consist of 6 features that play important any one the association rule mining directly. Thus we applied fuzzy classification first and then after fuzzification , genetic algorithm is applied to find the frequently associated itemsets.

With the help of the results, frequently associated itemsets can be used to find out the association among the different factors that can cause heart disease. So if a person is already satisfying 4 conditions then prior precautions can be taken to not to get affect with remaining two conditions that have high probability to appear and can cause heart disease. The purpose of this work to find the closely associated conditions that are very much frequent in the patients to have heart disease.

The experimental results for the above mentioned dataset used and fuzzy conditions and genetic implementation is given below:

1) ('Age = ', 'M', ' Blood Pressure = ', 'M', ' Cholesterol = ', 'M', ' Blood Sugar = ', '0', ' Heart Rate = ', 'M', ' ST Depression = ', 'N', ''), ' Support = ', 14, ' Support % = ', 4.682274247491639

2) ('Age = ', 'L', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'M', ' Blood Sugar = ', '0', ' Heart Rate = ', 'H', ' ST Depression = ', 'N', ''), ' Support = ', 15, ' Support % = ', 5.016722408026756

3) ('Age = ', 'M', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'L', ' Blood Sugar = ', '0', ' Heart Rate = ', 'L', ' ST Depression = ', 'SA', ''), ' Support = ', 5, ' Support % = ', 1.6722408026755853

4) ('Age = ', 'L', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'VH', ' Blood Sugar = ', '0', ' Heart Rate = ', 'M', ' ST Depression = ', 'N', ''), ' Support = ', 5, ' Support % = ', 1.6722408026755853

5) ('Age = ', 'L', ' Blood Pressure = ', 'M', ' Cholesterol = ', 'M', ' Blood Sugar = ', '0', ' Heart Rate = ', 'H', ' ST Depression = ', 'N', ''), ' Support = ', 6, ' Support % = ', 2.0066889632107023

6) ('Age = ', 'L', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'L', ' Blood Sugar = ', '0', ' Heart Rate = ', 'L', ' ST Depression = ', 'H', ''), ' Support = ', 3, ' Support % = ', 1.0033444816053512

7) ('Age = ', 'H', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'M', ' Blood Sugar = ', '0', ' Heart Rate = ', 'M', ' ST Depression = ', 'N', ''), ' Support = ', 3, ' Support % = ', 1.0033444816053512

8) ('Age = ', 'M', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'M', ' Blood Sugar = ', '0', ' Heart Rate = ', 'L', ' ST Depression = ', 'SA', ''), ' Support = ', 4, ' Support % = ', 1.3377926421404682

9) ('Age = ', 'M', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'L', ' Blood Sugar = ', '0', ' Heart Rate = ', 'M', ' ST Depression = ', 'SA', ''), ' Support = ', 5, ' Support % = ', 1.6722408026755853

10) ('Age = ', 'M', ' Blood Pressure = ', 'L', ' Cholesterol = ', 'L', ' Blood Sugar = ', '0', ' Heart Rate = ', 'M', ' ST Depression = ', 'SA', ''), ' Support = ', 5, ' Support % = ', 1.6722408026755853

VI. CONCLUSION AND FUTURE WORK

In this paper authors have used the fuzzy logic to categorize the numerical values with membership values ranges from 0 to 1 and enormous robustness of Genetic algorithm to get the list of frequent itemset. Frequent itemsets can be used further to get the association rules. By using different membership functions for different factors , use of apriori algorithm [10] to get frequent itemset will cost very high. This paper can be extend to generate association rule from the generated frequent itemsets in an efficient manner. So definitely this Hybrid system will be better and reduce the overall complexity.

REFERENCES

[1] Agrawal, R., Imielinski, T. and Swami, A. “Mining Association Rules between Sets of Items in Large Database”. Proceedings of the ACM SIGMOD conference on management of data, Washington, D.C, May 26-28, 1993.

[2] M. Kowsigan, A.Christy Jebamalar, S. shobika, R. Roshini, A. saravanan, “ Heart Disease Pridiction by analysing various parameters using fuzzy logic”, Biotechnol www.pjbt.org, Vol(2) 157-161, 2017.

[3] Manish saggar, Ashish Kumar Agrawal, Abhimanyu Lad, “ Optimization of Association Rule Mining using Improved Genetic Algorithm”, IEEE, 2004, pp 3725-3729.

[4] Savasere A., E. Omieccinski and S. Navathe, “ An efficient algorithm for mining association rules in large database”. Proceeding of the 21st International Conference on very large database, Zurich, Switzerland, Sept 11 – 15 ,1995, pp 432-443

[5] Weining Zhang, “ Mining Fuzzy quantitative Association Rules”, ICTAI '99 Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, IEEE,8-10 November, 1999.

[6] Fan Jiancong, Liang Yongquan, Ruan Jiuhong, “An Evolutionary Mining Model in Incremental Data Mining”, Fifth International Conference on Natural Computation, IEEE, 2009, pp: 114-118.

[7] Anil Vasoya and Nitin koli, “Mining of Association Rules on Large Database Using Distributed and Parallel Computing”, Proceedings of International Conference on Communication, Computing and Virtualization (ICCCV) 2016, Volume 79, 2016, Pages 221–230 (Procedia Computer Science).

[8] Yongfu Wang, Hong Zhao, Jiren Liu, “Fuzzy Modeling method based on Data Mining”, Proceedings of the 7th world congress of intelligent control and automation, June 2008, China.

[9] Huang Wei, “ Study on a data warehouse mining oriented fuzzy association rule mining algorithm”, fifth international conference on Intelligent System Design and Engineering applications, 2014.

[10] J. Pei, J. Han, and L.V.S. Lakshmanan, “Mining frequent itemsets with convertible constraints”. In Proc. ICDE 2001, pp. 433–442.

[11] Brin, S., Motwani, R., and Silverstein, C. “Beyond market baskets: Generalizing association rules to correlations”. SIGMOD 26[2], 265-276. 1997.

[12] D.Subalakshmi, C.Rajagopal, G.Santoshkumar, “Fuzzy Expert System For Heart Diseae Diagnosis”, International Journal of Pure and Applied Mathematics Volume 116 No. 21 2017, 349-353 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) url: http://www.ijpam.eu Special Issue.

[13] Rajdeep Kaur Aulakh, “Optimized Association Rule Mining with Maximum Constraints using Genetic Algorithm”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 3 Issue IV ISSN: 2321-9653, April 2015.

[14] Ayush Kumar Agrawal, Rohit Miri, S.R.Tandan, “A Review on Fuzzy Mining Association Rule Techniques”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue V ISSN: 2321-9653, May 2016.

[15] Basheer Mohamad Al-Maqaleh, “Discovering Interesting Association Rules: A Multi-objective Genetic Algorithm Approach”, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 5– No.3, February 2013.

[16] Mimanshu Gupta, Beerendra Kumar, Rohit Miri, “A Survey Paper on Association Rule Mining using Fuzzy Logic”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 3 Issue V ISSN: 2321-9653, May 2015.