K. Rajeswari, Mahadev Shindalkar, Nikhil Thorawade, Pranay Bhandari / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
www.ijera.com Vol. 3, Issue 4, Jul-Aug 2013, pp.132-136

DSS Using Apriori Algorithm, Genetic Algorithm And Fuzzy Logic

K. Rajeswari Professor [1], Mahadev Shindalkar Student [2], Nikhil Thorawade Student [3], Pranay Bhandari Student [4]

[1, 2, 3, 4](Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune-44, Maharashtra, India)

ABSTRACT

DSS is an interactive software system which assists the user (Doctors, Stock Analyzer etc.) to make appropriate and quick decisions in the given context. The volume of data, in fields like medicine or industry, is large and beyond human capacity to extract valuable information. Apriori algorithm along with genetic algorithm and fuzzy logic provides a way to obtain interesting correlations and patterns from datasets. It also eliminates human error and provides high accuracy of results. Therefore the DSS helps to make quick and accurate decisions Experiments were conducted on Pima Indian Diabetes data set and results are found to be interesting.

Keywords **-** Apriori, DSS, Fuzzy Logic, Genetic.

I. INTRODUCTION

In today's world people are flooded by data like scientific data, medical data, financial data and marketing data. However the Human capacity is limited and a common human cannot deal with such massive amount of data to extract the information. The solution is to develop automatic techniques and systems – to analyze the data, to classify the data, to summarize it, to discover and characterize trends in it [11].

Data mining is the extraction of interesting relations and patterns hidden in the datasets. It combines database technologies, statistical analysis and machine learning[1].Today data mining techniques are innovatively utilized in numerous fields like industry, Commerce and Medicine. The data results generated by data mining techniques is highly priced by the professionals [2].For Instance in this paper, medical data is used as a context for mining frequent patterns, associations using the apriori algorithm. Further, Genetic algorithm and fuzzy logic will be utilized for decision making. Here we propose a knowledge based DSS for use in data processing. In general the knowledge based DSS consists of 3 fundamental parts – Database or dataset, inference rules and user interface for the user to understand the results.

The classical apriori algorithm is one of the most conventional and significant data mining algorithms which are used for identifying the relationships among attributes of transactions using binary values.

However, mostly the real life applications have transactions with quantitative values which are not binary.

1.1 Extraction of Interesting association patterns using Apriori

The data to be processed by the DSS has to be free of noises, but it can exhibit uncertainty associated with it. As a result, training the DSS directly with pre-processed data leads to incorrect inference. Thus, to obtain interesting patterns from the data, we have to apply data mining techniques on the pre-processed data and then utilise the results as base data for the DSS. In this paper we use classical and improved apriori to discover association rules hidden in the data. The association rules are the inter-relations among the attributes based on the item-sets defined in the dataset. Apriori algorithm [3] is the most significant algorithm when it comes to mining association rules.

Apriori key concepts:

1) Association rules – There are two types or categories of association rules :
   a) Boolean association rule
   b) Quantitative association rule
2) Minimum Support – It's the percentage of task-relevant data transactions for which the pattern is true.
3) Minimum Confidence threshold – It's the measure of certainty associated with the pattern.
4) Itemset - K-items in a set of items is called K-item set.
5) Frequent item-sets – Item sets with occurrence frequency greater than or equal to the minimum support are called frequent item-set (Denoted by L_i for i^{th} item set).
6) Strong rules – Association rules which that satisfy both min_support and min_confidence are called strong rules.
7) Apriori property – Subset of a frequent itemset is always frequent.

8) Join operation – To generate new combinations of itemsets following the apriori property .

Let the set of frequent itemset be F_k and the set of candidate itemset be C_k.
Steps in apriori:-

1) Generate the C_{k+1} candidate set for generation of F_{k+1} frequent itemset.
2) Scan the database and determine the value of support of each candidate of frequent itemset.
3) Then sort out those itemsets which have support value more than min_support. This will generate the Fk+1.

Pseudo- Code is explained as follows.
F_i = (Frequent itemsets with cardinality i);
for (k = 1; $F_k \neq \varphi$; k + +) do begin
C_{k+1} = *Apriori_Candidate_gen* (F_k); // new candidates
for every transactions $t \in$ database do begin' ( )
C_t = *subset* (C_{k+1}, t) ; // Candidates contained in t
for all candidate '
$c \in C_t$ do
c.count + +
end
F_{k+1} = {C$\in$ C_{k+1}| *c.count* $\geq$ *minimum support*
end
end
Answer: $U_k F_k$

The Apriori-gen function in above pseudo-code generates the C_{k+1} candidate set from the F_k frequent iteset in two step process:

1) Join step – To find C_{k+1}-a set of candidate k-item sets, it is generated by joining L_k with itself.
2) Prune step – k-Itemsets which do not satisfy min_support value are removed from join set to generate the candidate itemset.

1.2 Genetic Algorithm

The main purpose of DSS is to predict the risk of any disease. Doctors and specialists have been doing this based on their knowledge, experience and medical reports. Over the course of time they have developed a grand edifice of knowledge that enables them to predict to varying extents. But there lies fundamental limits to our ability to predict. The solution to this is an optimizing tool that gives the best and optimized results. Genetic algorithm is a tool used for this purpose [5]. The idea is to evolve a population of solutions based on some function or condition, using operators inspired by genetic variation and natural selection.

Genetic algorithm facts are as follows:

- Genetic algorithms (GAs) were invented by John Holland in the 1960s and were developed by Holland and his students and colleagues at the University of Michigan in the 1960s and the 1970s.
- Heuristic Search Algorithms Method based on evolutionary ideas of natural selection and genetics
- Provides efficient, effective techniques for optimization
- Useful when search space very large or too complex for analytic treatment

Algorithm Key Concepts are as follows:

1. Individual - Any possible solution
2. Genes-Attributes of an individual
3. Population - Group of all *individuals*
4. Search Space - All possible solutions to the problem
5. Chromosome – (set of genes)Blueprint for an *individual*
6. Fitness function- A function that assigns a fitness value to an individual
7. Genetic Operators:
   - Reproduction(Selection)
   - Crossover(or Recombination)
   - Mutation-(Changing or Modifying)

The steps involved in determination of optimal set of attributes are:

- *Inputs:*
  1. Medical database
  2. Standard range of attributes(e.g. cholesterol, BP, etc.) from expertise
  3. Rules generated from Apriori
- *Initial Requirement:* Dataset needs to be pre-processed for better decision making.
- *Output:*
  Optimal values of attributes i.e. the best chromosome with highest fitness value
- *Size of Population:*
  Suppose no. of attributes is N. Therefore initial population will consist of N chromosomes.
  But total size of population will be 2N where next N will consist of newly generated population.
- *Chromosome initiation:* Chromosomes will consist of one distinct value for every attribute. Hence size of chromosomes will be N.(Here attribute is gene)
- *Fitness function:*
  In the proposed system, the fitness function calculates the optimal values of attributes based on three parameters:
  i. Actual value i.e. the gene value

ii. Deviation from normal values (the standard values obtained from expertise)
iii. Frequency of the value of an attribute

Steps of Fitness Calculation:

1. Check if gene value lies in the range given by Expert.
   If no, then calculate the deviation from normal value using:
   Deviation = gene[i] value – Normal value, 0<=i<N
   Else Deviation=0
2. Calculate fitness of a gene using formula:
   = (Frequency of occurrence of gene[i] * Deviation)/N

- *Crossover and Mutation:* Single-point crossover has been used in the proposed system. This crossover point is generated randomly. For mutation, swap method is used followed by fitness calculation. These operators generate a complete new set of population.

Algorithm:

1. Start- Generate initial population of n chromosomes
2. Evaluate- Calculate the fitness f(x) for each chromosome x in n
3. Iteratively for each consecutive pair of chromosomes
   Generate new population using: -
   - Selection (This operator selects chromosomes in the population for reproduction.)
   - Crossover (This operator randomly chooses a point of crossover and exchanges the sub sequences before and after that point between two chromosomes to create two offspring.
   - Mutation (This operator randomly flips some of the bits in a chromosome. This is most general method. Nevertheless other methods also exist.)
   - Calculate Fitness – A fitness function is defined according to the problem to select the better chromosomes for further consecutive generations.
4. Sort the chromosomes based on their fitness values to get topmost solution as the best solution
5. Steps 3, 4gives one generation. Repeat these steps until :
   5.1. No. of specified no. of generations reached OR
   5.2 Required best solution achieved
6. Stop- Return the best solution in current population

After m number of generations, the GA results with the final optimal set of 2N chromosomes. These optimal set of chromosomes determined by GA represent the high impact attribute values for diagnosis. Subsequently, the first chromosome from the optimal list is chosen and their corresponding values are recorded as optimal value for each attribute. These optimal values serve as values for linguistic variables used in the design of the Fuzzy membership function.

1.3 Decision module using Fuzzy Logic (FL)

The Fuzzy inference process consists of 5 fundamental steps:

a) Fuzzification of input variables
b) Application of AND or OR operators
c) Implication from antecedent to consequent
d) Aggregating the consequents over the generated rules
e) De-Fuzzification

Steps of Fuzzy inference process

1) Fuzzification of the inputs – Initially it is determined to what degree the input belongs to a particular membership function. Input given to the FL Toolbox is a limited numeric value in a given range. The output is always between 0 and 1 which is a fuzzy membership.
2) Applying the AND or OR operators – AND, OR are the fuzzy operators. Two or more membership values from fuzzified input variables are given as input to the fuzzy operator and a single truth values is obtained as the output.
3) Applying implication method – The weight associated with a rule ranges from 0 to 1 and is assigned to a number given by the antecedent.
4) Aggregate all outputs: Aggregation occurs immediately before De-fuzzification and occurs only once for every output variable.
5) Defuzzification: In de-fuzzification it takes a fuzzy set (the output of the aggregate step) as input and then, gives a single number as output.

II. LITERATURE REVIEW

Researchers in the past have proposed many DSSs to assist professionals like doctors, engineers, market analyzers, in decision making. But there has not been a highly successful implementation of any DSSs in the past. Our team has proposed a DSS which utilizes Apriori, Genetic algorithm and Fuzzy logic for decision making.

The design of DSS which use data mining and Artificial intelligence (AI) techniques for

decision making is one big research area. In this section we appreciate important contributions to this field from existing literature. GO Barnett, JJ Cimino (1987) proposed a computer based DSS called "DXplain"[8] which accepts a list of clinical manifestations and then proposes diagnostic hypotheses.

Latha and Subramanian (2008) proposed an approach based on CANFIS (Coactive Neuro-Fuzzy Inference System) to predict risk of Heart disease [9]. The CANFIS based system predicts diagnosis with the combination of neural network's adaptive capabilities and FL's qualitative approach which are later integrated with Genetic algorithm. The effectiveness of the CANFIS model has been proved by expressing its performance in terms training performances and classification accuracies.

Y Huang, R Desiraju (2000) have proposed a DSS for management of an Agile Supply Chain [7]. The system proposed has Client-Server architecture. The server side has the DSS database that interfaces with the model engine which analyses the data. The server also consists of a server manager which processes requests and information.

Shanthi et al. (2008) proposed a neuro-genetic approach of feature selection in classification of diseases [12]. Three-layered feed-forward neural network are used for examining the Candidate feature subsets. Their system predicts diagnosis of stroke diseases. The system has a multilayered perceptron (MLP) which automatically selects its inputs using GA for dimensionality reduction. Experimental results have proved their approach to be better in classification accuracy with fewer input features.

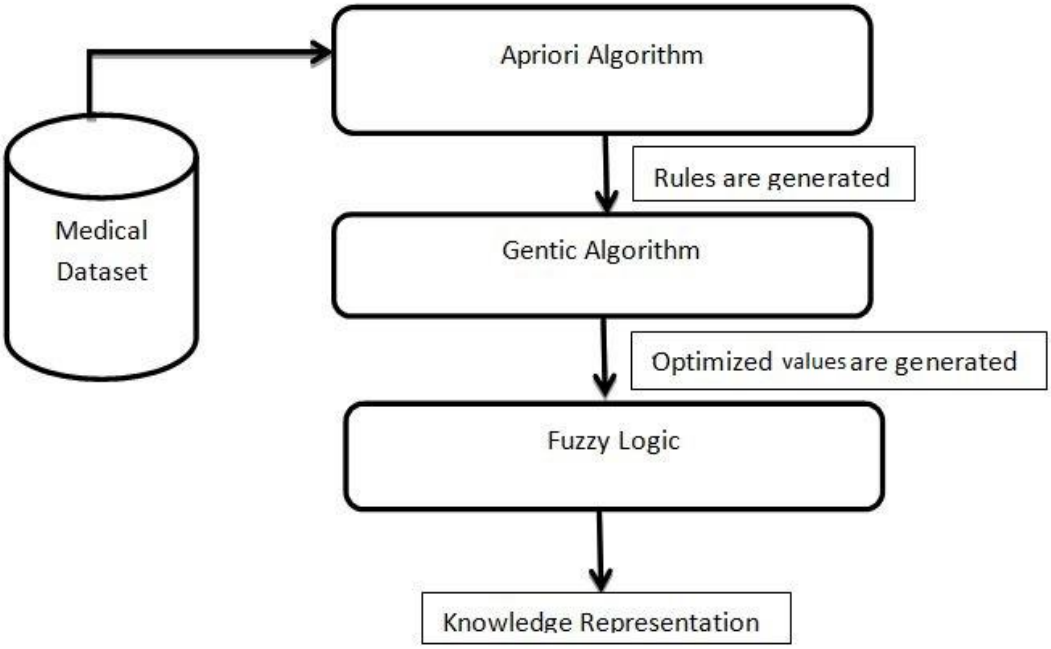

Fig.1. System Workflow

Lin et al. (2006) have innovatively proposed a web-based DSS[10] to analyse and evaluate the medical data of patients. The system also predicts suggestions on diagnosis. Their system deals with the characteristics of Lower Back Pain (LBP). The system uses a systematic method of evaluation which includes verifying knowledge base, modified Turing test for system validation and assessing the clinical efficacy involving 5 doctors and 180 real world cases gathered from various clinics.

III. METHODOLOGY

We have taken the standard benchmark dataset and processed it according to steps shown in the Fig.1.

1. Initially, the standard dataset is taken for evaluation by the system.
2. The dataset is pre-processed for conversion of quantitative values into relative values.
3. The pre-processed data is given to the Apriori algorithm for generation of associations and interesting correlations hidden in the data.
4. The results of apriori module are given to Genetic algorithm module to optimise the results i.e. to sort out the most interesting association rules from apriori results.
5. The optimised results from Genetic algorithm module are given to the Fuzzy logic module for generation of required knowledge.

IV. RESULTS AND DISCUSSION

Some Apriori Rules from Weka

1. slope=a5 303 ==> Cp=a5 303 conf:(1)
2. Cp=a5 303 ==> slope=a5 303 conf:(1)
3. fbs=a0 258 ==> Cp=a5 258 conf:(1)
4. fbs=a0 258 ==> slope=a5 258 conf:(1)
5. fbs=a0 slope=a5 258 ==> Cp=a5 258 conf:(1)
6. Cp=a5 fbs=a0 258 ==> slope=a5 258 conf:(1)
7. fbs=a0 258 ==> Cp=a5 slope=a5 258 conf:(1)
8. Sex=a5 206 ==> Cp=a5 206 conf:(1)
9. Sex=a5 206 ==> slope=a5 206 conf:(1)
10. Sex=a5 slope=a5 206 ==> Cp=a5 206 conf:(1)
11. Sex=a5 Cp=a5 206 ==> slope=a5 206 conf:(1)
12. Sex=a5 206 ==> Cp=a5 slope=a5 206 conf:(1)
13. exang=a0 204 ==> Cp=a5 204 conf:(1)
14. exang=a0 204 ==> slope=a5 204 conf:(1)
15. exang=a0 slope=a5 204 ==> Cp=a5 204 conf:(1)

These rules are given to genetic algorithm. It gives optimized rules.

Output of genetic algorithm:

The number of Genes per Chromosome is: 9
Genes: Nos_preg, plasma, diastolic, Triceps, serium, massIndex, Pedigree, Age, Class
The numbers of chromosomes are: 10
Chromosome 1 is :
5.0,141.0,70.0,35.0,230.0,30.0,5.0,25.0,5.0 : :260.0

Chromosome 1 represents optimized values. These optimized rules are given to fuzzy logic.

After drafting membership functions in Matlab, following results are obtained.

Fuzzy logic Output

5 130 71 32 10 41 2.5 60 -> Risk
5 142 60 10 20 32 1.4 23 -> Risk

Therefore, this process will help to predict the risk of diseases and making effective decisions.

V. CONCLUSION

In this paper we have proposed and developed a Decision support system which will assist doctors in processing medical data of patients in to predict risk of disease.

REFERENCES

[1] Thuraisingham, B., 2000. A primer for understanding and applying data mining, *IT professional.IEEE Computer Society*, pp. 28–31.

[2] Tang, T.I., Zheng, G., Huang, Y.,Shu, G., Wang, P., 2005. A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. IEMS 4 (1), 102–108.

[3] Agrawal, R. and Srikant, R. (1994) 'Fast algorithms for mining association rules', Proceedings *of the 20th International Conference on Very Large Data Bases*, pp.487–499, Santiago de Chile,

[4] Anooj, P. K. (2012). Clinical decision support system : Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University - Computer and Information Sciences*, *24*(1), 27–40. doi:10.1016/j.jksuci.2011.09.002

[5] Kuo, R. J., Chen, C. H., & Hwang, Y. C. (2001). An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artiÿcial neural network, *118*, 21–45.

[6] Rajeswari, K., & Vaithiyanathan, V. (2011). Heart disease diagnosis : an efficient decision support system based on fuzzy logic and genetic algorithm, *3*, 81–97.

[7] Huang, Ying, et al. "Decision support system for the management of an agile supply chain." U.S. Patent No. 6,151,582. 21 Nov. 2000.

[8] Barnett, G. O., Cimino, J. J., Hupp, J. A., & Hoffer, E. P. (1987). An evolving diagnostic decision-support system. *Jama*, *258*, 67-74.

[9] Latha, P. and Subramanian, R. (2008) 'Intelligent heart disease prediction system using CANFIS and genetic algorithm', *International Journal of Biological and Medical Sciences*, Vol. 3, No. 3.

[10] Lin, L., Hu, P.J-H. and Sheng O.R.L. (2006) 'A decision support system for lower back pain Diagnosis: uncertainty management and clinical evaluations', *Decision Support Systems*, Vol. 42, No. 2, pp.1152–1169.

[11] Han, Jiawei, and Micheline Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

[12] Shanthi, D., Sahoo, G. and Saravanan, N. (2008) 'Input feature selection using hybrid neurogenetic approach in the diagnosis of stroke disease', *International Journal of Computer Science and Network Security*, Vol. 8, No. 12, pp.99–107.