Optimization and Improvement of Association Rule Mining using Genetic Algorithm and Fuzzy Logic

Archana Gupta[a], Sanjeev Jain[b], Akhilesh Tiwari[a]

[a]*Madhav Institute of Technology and Science, Gwalior (M.P.), India*
[b]*Shri Mata Vaishno Devi University(SMVDU), Katra (J&K), India*

ARTICLE INFO

Article history:
Received 19 January 19
Received in revised form 30 January 19
Accepted 23 February 19

Keywords:
Association Rule Mining
Genetic Algorithm
Fuzzy Logic

ABSTRACT

In Association rule mining, association rules grows exponentially with the size of the frequent itemset. With so many association rules one is interested in the optimized rules to make decisions. With huge attribute's domain, the rules generated in the association rule mining are also huge. In this paper, the main focus is on how to optimize the association rules using Genetic Algorithm and How to improve the working of Association Rule Mining using Fuzzy Logic. With explosion of data, optimization of the results is as important as producing the results. Genetic Algorithm perform global search to find attribute interaction than the greedy rule induction algorithm often used in Data Mining. To show the use of fuzzy logic and Genetic Algorithm in Association Rule Mining, a system to predict the chance of having heart disease is considered in this paper.

Peer review under responsibility of International Conference on Sustainable Computing in Science, Technology and Management.

1. 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 (A Agrawal, R., Imielinski, T. and Swami) 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 traditional approach used in association rule mining, Genetic algorithm and Fuzzy set theory, section 3 gives the hybrid system used in association rule mining using genetic algorithm and fuzzy set theory for the chance of having heart disease, finally, the system will be concluded in section 4.

2. Basic Concepts

2.1. Association rule mining(ARM)

Association rule learning is a rule-based machine learning method for discovering interesting association among variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Initially Market basket analysis [A Agrawal, R., Imielinski, T. and Swami] 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 {bread, butter}->{milk} found in the sales data of a supermarket

would indicate that if a customer buys bread and butter together, they are likely to also buy milk. 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:

2.1.1. 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.

2.1.2. 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)

2.2. Genetic Algorithm

Genetic Algorithm (GA) [Fan Jiancong, Liang Yongquan, Ruan Jiuhong] 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. Implementation of GA makes use of bio-inspired operators such as selection, crossover and mutation. In a genetic algorithm, a population of candidate solutions (called individuals) to an optimization problem is evolved toward better solutions at every step. Each candidate solution has a set of properties i.e. its chromosomes which can be mutated and altered through the genetic algorithm. Individuals are normally represented in binary strings of 0s and 1s.

To implement genetic algorithm, there are following two requirements:

- a genetic representation of the solution/individual,
- a fitness function to evaluate the sustainability of solution/individual.

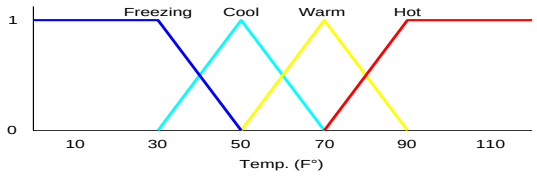
The evolution usually starts with the initial population that is created by randomly generated individuals [Manish saggar, Ashish Kumar Agrawal, Abhimanyu Lad]. It is an iterative process, apply GA operators - selection crossover and mutation with the population in each iteration called a generation. In each generation, the fitness of every individual in the population is evaluated based on the fitness function. Fitness value is usually the value of the objective function in the optimization problem being solved. The more fit individual is randomly selected in next generation (Kumar, Jain & Sharma, 2018).

2.3. 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 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 is given. It is employed to handle the concept of partial truth [Huang Wei], where the truth value may range between completely true and completely false.

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[M. Kowsigan, A.Christy Jebamalar, S. shobika, R. Roshini].

3. Hybrid System

In this section, Hybrid system is generated in the field of medical science to predict that what are deifferent factors that increase the chance of having heart disease. If strong association among the factors are known then a pateint can take the precaution to control that factor. So that the chance of have disease will get reduce. 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

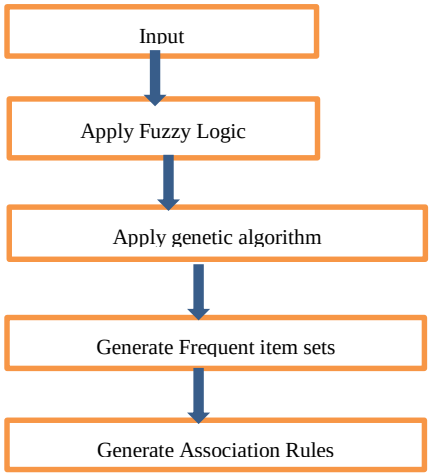


Fig 1. Hybrid system to generate Association rules

In second 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".

In the third 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.

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

In the fifth phase Association rules will get generate from the frequent itemset.

Following are the factors that would help to predict heart disease in an individual [M. Kowsigan, A.Christy Jebamalar, S. shobika, R. Roshini] :

- High blood pressure (hypertension)
- High LDL cholesterol (Bad Cholesterol)
- High HDL cholesterol (Good Cholesterol)
- High Triglycerides
- High Blood Glucose level
- Family history of premature heart disease
- Cigarette smoking
- Physical inactivity

Out of the given 8 factors, 5 factors will be used to predict the chance of having heart disease. The membership functions of the different linguistic variable used for these factors are given in subsequent section.

Blood Pressure [M. Kowsigan, A.Christy Jebamalar, S. shobika, R. Roshini]:

In blood pressure there are two categories 1. Systolic pressure and 2. diastolic pressure. Used three membership function to divide the range of blood pressures into three category of Normal, at Risk and High.

The range of the membership function for systolic pressure is given in table 1.

Table 1: Systolic pressure

Sr. No.	Fuzzy Set	Range	Membership function used
1	Normal	Below 120mmHg	Negative ramp
2	At Risk	120-139 mmHg	trapezoidal
3	High	140mmHg and above	Positive ramp

The range of the membership function for diastolic pressure is given in table 1.

Table 2 : Diastolic pressure

Sr. No.	Fuzzy Set	Range	Membership function used
1	Normal	Below 80mmHg	Negative ramp
2	At Risk	80-89 mmHg	Trapezoidal
3	High	90mmHg and above	Positive ramp

Total Cholesterol:

Total Cholesterol in body depends on LDL Cholesterol, HDL Cholesterol and triglycerides. It can be given as

Total Cholesterol= HDL+LDL+0.2*triglycerides

Different membership functions used for LDL Cholesterol, HDL Cholesterol and triglycerides are given in Table 3,4,5.

Table 3 : Low Density Lipoprotein Cholesterol

Sr. No.	Fuzzy set	Range	Membership function
1	Ideal	Below 100	Negative ramp
2	Close to ideal	100 to 129	Trapezoidal
3	Nearly high	130 to 159	Trapezoidal
4	High	160 to 189	Trapezoidal
5	Very high	190 and above	Positive ramp

Table 4 : High Density Lipoprotein Cholesterol

Sr. No.	Fuzzy set	Range	Membership function
1	Low	Below 40	Negative ramp
2	Normal	40 to 59	Trapezoidal
3	Best	Above 60	Positive ramp

Table 5 : Triglycerides

Sr. No.	Fuzzy set	Range	Membership function
1	Ideal	Below 150	Negative ramp
2	Border line	150 to 199	Trapezoidal
3	High	200 to 499	Trapezoidal
4	Very High	Above 500	Positive ramp

Blood Sugar level[M. Kowsigan, A.Christy Jebamalar, S. shobika, R. Roshini]:

Blood sugar level is given by three membership functions as Normal, High, Very High.

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 examples of the rule base [M. Kowsigan, A.Christy Jebamalar, S. shobika, R. Roshini] to predict the chance of having the heart disease.

Rule1 : if low density lipids is nearly high and high density lipid is low and triglycerides is high and systolic is low and diastolic is low then there is chance of heart disease to occur.

Rule 2: if low density lipids is high and high density lipid is low and triglycerides is border line and systolic is at risk and diastolic is low then there is chance of heart disease to occur.

Rule 3: if low density lipids is very high and high density lipid is moderate and triglycerides is nearly high and systolic is low and diastolic is high then there is chance of heart disease to occur.

Rule 4: if low density lipids is very low and high density lipid is low and triglycerides is very high and systolic is high and diastolic is very high then there is chance of heart disease to occur.

In these rules given in [M. Kowsigan, A.Christy Jebamalar, S. shobika, R. Roshini], here one more factor is added in chromosome, that is blood sugar level. If blood sugar level is high then the chance of having heart disease is usually there. But if blood sugar level is normal then other factors will over take the inference.

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:

- How to represent the chromosomes
- How to define the fitness function

Chromosome representation:

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

00 00 000 00 00 00

1 2 3 4 5 6

1. Systolic
2. Diastolic
3. LDL
4. HDL
5. Triglycerides
6. Blood sugar level

Chromosome corresponding to rule 1 is given as: 00 00 011 00 10

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.

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 met stopping criteria. Finally, a set of frequent itemsets can generate from final population, which can be further used to generate association rules.

4. Conclusion:

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 [J. Pei, J. Han, and L.V.S. Lakshmanan] to get frequent itemset will cost very high. So definitely this Hybrid system will be better and reduce the overall complexity.

REFERENCES

A 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.

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).

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

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.

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.

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

Kumar, S., Jain, S., & Sharma, H. (2018). Genetic Algorithms.In Advances in Swarm Intelligence for Optimizing Problems in Computer Science, pp. 27-52, Chapman and Hall/CRC.

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.

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

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

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.

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.