computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

OPTIMIZATION OF ASSOCIATION RULES FOR TUBERCULOSIS USING GENETIC ALGORITHM

Asha T. [1], S. Natarajan [2], K.N.B. Murthy [2]

[1] Bangalore Institute of Technology, India, asha.masthi@gmail.com
[2] PES Institute of Technology, India, natarajan@pes.edu, principal@pes.edu

Abstract: *Tuberculosis (TB) is a disease caused by bacteria called Mycobacterium Tuberculosis which usually spreads through the air and attacks low immune bodies. Human Immuno deficiency Virus (HIV) patients are more likely to be attacked by TB. It is an important health problem around the world including India. Association Rule Mining is the process of discovering interesting and unexpected rules from large sets of data. This approach results in huge quantity of rules where some are interesting and others are repetitive. It also limits the quality of rules to only two measures support and confidence. In this paper we try to optimize the rules generated by Association Rule Mining for Tuberculosis using Genetic Algorithm. Our approach is to extract only a small set of high quality Tuberculosis rules among the larger set using Genetic Algorithm. In the current approach datatypes such as discrete, continuous and categorical items have been handled. The proposed experimental result includes a small set of converged TB rules that helps doctors in their diagnosis decisions. The main motivation for using Genetic Algorithms in the discovery of high-level prediction rules is that they are robust, use adaptive search techniques that perform a global search on the solution space and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining.*

Keywords: *Tuberculosis, Data Mining, Diagnosis, Association Rules, Optimization, Genetic Algorithm.*

1. INTRODUCTION

Large amounts of data have been gathered routinely in the course of day-to-day management in business, administration, medicine, banking, the delivery of social and health services, environmental protection, security and in politics. Such data is primarily used for accounting and for management of the customer base. Typically, management data sets are very large and constantly growing but contain a large number of complex features. While these data sets reflect properties of the managed subjects and relations, and are thus potentially of some use to their owner, they often have relatively low information density. One requires robust, simple and computationally efficient tools to extract information from such data sets. The development and understanding of such tools is the core business of data mining. These tools are based on ideas from computer science, mathematics and statistics. Mining useful information and helpful knowledge from these large databases has thus evolved into an important research area called Data Mining.

Association Rule Mining is an important problem in the rapidly growing field called data mining and Knowledge Discovery in Databases (KDD) . The task of ARM is to mine a set of highly correlated attributes/features shared among a large number of records in a given database. For example, consider the sales database of a bookstore, where the records represent customers and the attributes represent books. The mined patterns are the set of books most frequently bought together by the customer. An example could be that, 60% of the people who buy Design and Analysis of Algorithms also buy Data Structure. The store can use this knowledge for promotions, self-placement etc. There are many application areas for Association Rule Mining techniques, which include catalogue design, store layout, customer segmentation, telecommunication alarm, diagnosis and so on.

In recent years there is an explosive growth of bio-medical data, ranging from those collected in pharmaceutical studies and cancer therapy investigations to those identified in genomics and proteonomics research. The digitization of critical medical information such as laboratory reports, patient records, research papers, and anatomic images has also resulted in large amounts of patient care data. These data need to be effectively

organized and analyzed in order to be useful. Therefore Data Mining is becoming an increasingly important tool which transforms large amount of these medical data into information.

India has the world's highest burden of TB with million estimated incident cases per year. It also ranks [1] among the world's highest HIV burden with an estimated 2.3 million persons living with HIV/AIDS. Tuberculosis is much more likely to be a fatal disease among HIV-infected persons than persons without HIV infection. It is a disease caused by mycobacterium which can affect virtually all organs, not sparing even the relatively inaccessible sites. The microorganisms usually enter the body by inhalation through the lungs. They spread from the initial location in the lungs to other parts of the body via the blood stream. They present a diagnostic dilemma even for physicians with a great deal of experience in this disease.

2. RELATED WORK

The Association Rule mining problem was first introduced in 1993 by Agrawal et al. [2, 3]. Agrawal et al. developed the Apriori algorithm for solving the Association Rule Mining problem. The Apriori-based algorithms work in two phases. The first phase is for frequent itemset generation where we find all sets of items (itemsets) whose support is greater than the user-specified minimum support. Such itemsets are called *frequent itemsets*. In the second phase we use the frequent itemsets to generate the desired rules using minimum confidence. The general idea is that if, say ABCD and AB are frequent itemsets, then we can determine if the rule AB=>CD. Support(X) is the fractions of the transactions in the database containing X and confidence is defined as the ratio support(X and Y) / support(X).Since the processing of the Apriori algorithm requires plenty of time, its computational efficiency is a very important issue. In order to improve the efficiency of Apriori, many researchers have proposed modified Association rule-related algorithms.

Savasere et al. [4] introduced a partition algorithm for mining Association rules that is fundamentally different from the classic algorithm. The algorithm logically divides the database into a number of nonoverlapping partitions, which can be held in the main memory. The partitions are considered one at a time and all large itemsets are generated for that partition. These large itemsets are further merged to create a set of all potential large itemsets. Then these itemsets are generated. Park et al. proposed the Direct Hashing and Pruning (DHP) algorithm in 1995. DHP can be derived from Apriori by introducing additional control. It makes use of an additional hash table that aims at limiting the generation of candidates as much as possible.

Toivonen proposed the sampling algorithm in 1996. The sampling algorithm applies the level-wise method to the sample, along with a lower minimal support threshold, to mine the superset of a large itemset. The method produces exact Association rules, but in some cases it does not generate all the Association rules, that is, some missing Association rules might exist [5]. The dynamic itemset counting (DIC) algorithm was proposed by Brin et al. in 1997. One of the main design motivations was to limit the total number of passes performed over databases. DIC partitions a database into several blocks marked by start points and repeatedly scans the database [6]. In contrast to Apriori, DIC can add new candidate itemsets at any start point, instead of just at the beginning of a new database scan.The Pincer-Search algorithm was proposed by Lin et al. in 1998 and can efficiently discover the maximum frequent set. The Pincer-Search algorithm combines both the bottom-up and top-down directions [7]. In 2001, Yang et al. proposed the efficient hash-based method for discovering maximal frequent Itemsets (HMFS). The HMFS method combines the advantages of both the DHP and the Pincer-Search algorithm [8].

ARM with Apriori has been used by many researchers for the prediction of different diseases. Predicting heart disease has been automated by Association rules. An algorithm that uses search constraints to reduce the number of rules was proposed which searches for Association rules on a training set, and finally validates them on an independent test set. The medical significance of discovered rules is evaluated with support, confidence, and lift [24, 25, 37]. Association rules describe what drugs are frequently coprescribed with antacids [38]. Association rules are applied for the analysis of human sleep data using polysomnographic readings. The authors introduced a specialized Association Rule Mining technique that can extract patterns from complex sleep data comprising polysomnographic recordings, clinical summaries, and sleep questionnaire responses. The rules mined can describe associations among temporally annotated events and questionnaire or summary data; e.g., the likelihood that occurrences of a rapid eye movement (REM) sleep stage during the second 100 sleep epochs of the night is associated with moderate caffeine intake [23].

Though various implementations of Apriori discussed above can be used for deriving rules, still this approach results in huge quantity of rules where some are interesting and others are

repetitive. It also limits the quality of rules to only two measures, support and confidence. Hence Optimization algorithms have been widely used on Association Rule Mining to better improve the rules [9, 39, 40, 41, 42]. Genetic Algorithm is one such optimization technique.

3. GENETIC ALGORITHM (GA)

GAs represent a new programming paradigm that tries to mimic the process of natural evolution to solve computing and optimization problems. It is a stochastic global search method that operates on a population of potential solutions applying the principle of survival of the fittest to produce (hopefully) better and better approximations to a solution [18]. The mining of large data sets by Genetic Algorithms has only recently become practical due to the availability of affordable, high speed computers. GAs in data mining may also be used to evaluate the fitness of other algorithms.

Individuals, or current approximations, are encoded as strings called *chromosomes*, composed over some alphabet(s), so that the *genotypes* (chromosome values) are uniquely mapped onto the decision variable (*phenotypic*) domain. The most commonly used representation in GAs is the binary alphabet {0, 1} although other representations can be used, e.g. ternary, integer, real-valued etc.

Having decoded the chromosome representation into the decision variable domain, it is possible to assess the performance, or *fitness*, of individual members of a population. This is done through an objective function that characterizes an individual's performance in the problem domain. In the natural world, this would be an individual's ability to survive in its present environment. Thus, the objective function establishes the basis for selection of pairs of individuals that will be mated together during reproduction. The different genetic operators are:

- SELECTION deals with the probabilistic survival of the fittest, in that more fit chromosome are chosen to survive. Fitness is a comparable measure of how well a chromosome solves the problem at hand.
- CROSSOVER takes individual chromosomes from population P and combines them to form new ones.
- MUTATION alters the new solutions so as to add stochasticity in the search for better solutions [18].

Fig.1 explains the flow of how genetic algorithm works.

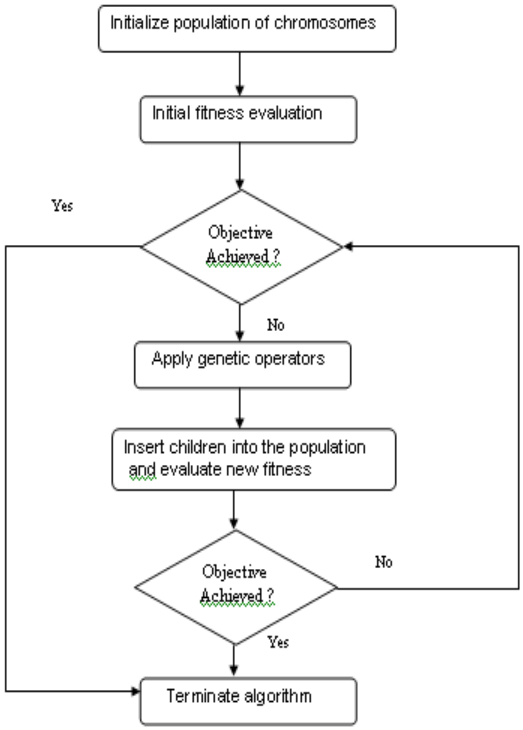

Fig. 1 – Schematic representation of a genetic algorithm

R. J. Kuo et al. [9] proposed the use of particle swarm optimization to search for the optimum fitness value of each particle and then finds corresponding support and confidence as minimal threshold values after the data are transformed into binary values. Saggar et al. [10] proposed an approach concentrating on optimizing the rules generated using genetic algorithms. The most important aspect of their approach is that it can predict the rule that contains negative attributes. Halina Kwasnicka et al. presented a method of Association rules discovering from medical data using evolutionary approach EGAR [11].

Ashish Ghosh and Bhabesh Nath [12] used Pareto based genetic algorithm to extract some useful and interesting rules from any market-basket type database. Sufal Das and Banani Saha have[13] considered four data qualities like accuracy, comprehensibility, interestingness and completeness to develop Multi-objective Genetic Algorithm (GA) utilizing linkage between feature selection and Association rule. In addition to this Peter P. Wakabi et al. [14] also presented a Pareto-based multiobjective evolutionary algorithm rule mining method based on genetic algorithms.

Orhan Er. and Temuritus [15] presented a study on Tuberculosis diagnosis, carried out with the help of Multilayer Neural Networks (MLNNs). For this purpose, an MLNN with two hidden layers and a genetic algorithm for training algorithm has been used. Data mining approach was adopted to classify genotype of mycobacterium tuberculosis using c4.5

algorithm [16]. Rethabile Khutlang et.al. presented methods for the automated identification of Mycobacterium Tuberculosis in images of Ziehl–Neelsen (ZN) stained sputum smears obtained using a bright-field microscope. They segment candidate bacillus objects using a combination of two-class pixel classifiers [17]. It can be observed from the literature survey that most of the work on TB has been carried out using Neural Network for classification. But Neural Networks lack knowledge representation. Further no work has been published on TB using ARM and optimization techniques. The rules extracted from TB help physician in diagnosing the disease by finding out the association between one symptom with the other. This explored aspect makes the study an important one.

Stochastic Universal Sampling (SUS) is one among the selection technique. The selection function is one that chooses parents for the next generation based on their scaled values from the fitness scaling function. An individual can be selected more than once as a parent, in which case it contributes its genes to more than one child. SUS is a single-phase sampling algorithm with minimum spread and zero bias. Instead of the single selection pointer employed in roulette wheel methods, SUS uses *N* equally spaced pointers, where *N* is the number of selections required. The population is shuffled randomly and a single random number in the range [0 *Sum*/*N*] is generated, *ptr*. The *N* individuals are then chosen by generating the *N* pointers spaced by 1, [*ptr*, *ptr*+1, ..., *ptr*+*N*-1], and selecting the individuals whose fitnesses span the positions of the pointers. In addition, as individuals are selected entirely on their position in the population, SUS has zero bias. The roulette wheel selection methods can all be implemented as O(NlogN) although SUS is a simpler algorithm and has time complexity O(N). GA differs substantially from more traditional search and optimization methods. The four most significant differences are:

- GAs searches a population of points in parallel, not a single point.
- GAs do not require derivative information or other auxiliary knowledge; only the objective function and corresponding fitness levels influence the directions of search.
- GAs use probabilistic transition rules, not deterministic ones.
- GAs work on an encoding of the parameter set rather than the parameter set itself (except in where real-valued individuals are used).

4. DATA SOURCE

The medical dataset we are using includes 700 real records of patients suffering from TB obtained from a city hospital. The entire dataset is put in one file having many records. Each record corresponds to most relevant information of one patient. Initial queries by doctor as symptoms and some required test details of patients have been considered as main attributes. Totally there are 12 attributes (symptoms) and last attribute indicates the type of TB. The symptoms of each patient such as age, chronic cough (weeks), loss of weight, intermittent fever (days), night sweats, Sputum, Blood cough, chest pain, HIV, radiographic findings, wheezing and TB type are considered as attributes. Table 1 shows names of 12 attributes considered along with their Data Types (DT). Type N-indicates numerical and C is categorical.

Table 1. List of Attributes and their Datatypes

No	Name	DT
1	Age	N
2	chroniccough(weeks)	N
3	weightloss	C
4	intermittentfever(days)	N
5	nightsweats	C
6	Bloodcough	C
7	chestpain	C
8	Diabetesmellitus	C
9	HIV	C
10	Radiographicfindings	C
11	Sputum	C
12	wheezing	C
13	TBtype	C

5. METHODOLOGY

In this section, we explain our proposed method where we first generate Tuberculosis Association rules and then apply Genetic Algorithm to optimize the rules.

Proposed method

1] Generate AR using Apriori

- Load the original Tuberculosis dataset
- Preprocess the dataset by discretization and Normalization
- Generate rules by applying Apriori on preprocessed range data

2] Optimize the output of step [1] using genetic algorithm

- Load the Tuberculosis Association rule database from step 1
- Encode each rule as an individual chromosome
- Initialize the population of individuals
- Calculate support, confidence and lift of each individual by scanning original TB data

- Evaluate the fitness of each individual

Repeat

- Select the best parents
- Crossover and mutate to generate offsprings
- Calculate the fitness of new offsprings
- Replace unfit individuals in the population with new ones

UNTIL stopping criteria is met.

5.1 TUBERCULOSIS ASSOCIATION RULE GENERATION USING APRIORI

The original raw TB data is first preprocessed by discretizing and normalizing numerical and categorical attributes respectively. Discretization techniques can be used to reduce the values of given continuous numerical attributes by dividing the range of attribute into intervals. Suppose if age attribute takes on the values from 1 to 100, it is treated as a continuous item and discretization reduces this into four intervals: 0<=age<=25, 26<=age<=50, 51<=age<=75, 76<=age<=100 and assigns to each interval a unique column number 1, 2, 3, 4 respectively. Normalization converts values associated with categorical data items so that they correspond to unique integer labels. A categorical data item `colour` which can have a value taken from the set `{blue, green}` can be normalised by assigning the column numbers 5, 6 to the possible values. Fig. 2 displays the sample preprocessed output.

Then Apriori algorithm [2, 3] is applied on the preprocessed output in obtaining the rules. Tuberculosis rules are generated by first extracting frequent itemsets from TB data. Different support and confidence values are used to generate frequent sets of TB. In the second step several medically important Association Rules are obtained with the help of rule generation concept. With 60% support and 80% confidence, we could extract around 227 rules. Following Fig. 3 provides sample output of the generated rules which are in preprocessed value format and can be converted back to their original attribute values. The last value in each row indicates the rule's confidence.

5.2 TB ASSOCIATION RULES ENCODING

For any individual representation basically there are two approaches to represent the rules, named as Pittsburgh and Michigan. In the Pittsburgh approach each chromosomes represents a set of rules and this approach is more suitable for classification rule mining; as we do not have to decode the consequent part and the length of the chromosome limits the number of rules generated. The other approach is called Michigan approach where each chromosome represents a separate rule. A modified approach is currently proposed by Ghosh et al. [12]. In this approach each attribute is tagged with two bits. If these two bits are 00 then attribute next to these two bits appears in the antecedent part and if it is 11 the attribute appears in the consequent part. The other two combinations, 01 and 10 will indicate the absence of the attributes in either of these parts. For example if there are A to F attributes and any rule is of the form as in equation (1)

A B -> E F is encoded as 00A 00B 01C 01D 11E 11F (1)

There is no restriction on the number of consequents. Binary encoding has been used here.

```
1 5 9 11 18 20 21 24 26 91 94 95
1 5 9 11 18 20 22 24 25 91 94 95
2 5 10 11 18 20 22 24 25 91 94 95
3 6 9 11 18 20 21 24 27 91 94 95
3 5 9 11 15 20 22 24 28 91 93 95
2 5 9 11 15 20 22 23 25 92 94 96
2 5 9 11 18 19 22 24 25 91 93 95
3 5 10 11 18 20 22 24 25 91 94 95
1 5 9 11 18 20 22 23 25 91 93 96
2 5 10 12 15 20 22 23 25 91 94 96
2 5 9 12 18 20 22 23 25 91 93 96
2 5 10 11 15 20 22 23 25 92 94 96
3 5 10 11 18 20 21 23 29 91 94 96
1 5 10 11 18 20 21 23 25 91 94 96
3 5 9 11 18 20 21 24 30 91 93 95
2 5 9 11 16 20 21 23 25 91 94 96
3 5 10 11 18 20 22 24 31 92 94 95
1 5 10 11 16 20 22 24 32 91 94 95
3 5 9 11 15 19 22 24 90 91 94 95
3 5 10 11 15 20 22 24 90 91 93 95
2 5 10 11 18 20 22 24 33 91 94 95
2 5 9 11 15 20 21 24 34 91 93 95
```

Fig. 2 – Sample preprocessed output

Rule representation for attribute is different in our approach where we consider the discretized range of values per attribute. Numerical attribute takes four ranges where each is given a different column number and categorical with two column numbers as described above. Hence numerical attributes require 2 bits and categorical attribute require1 bit for the representation along with tag bits. Though the range of different attributes are different and column numbers of each increases for the next attribute, it is seen that each attribute is reduced to 2 bits for numerical attribute with four status values 00 for first column number, 01-second, 10-third, 11-fourth column. Similarly 2 bits are used for categorical attribute with 0-first number and 1-second. For eg: if Tuberculosis rule from Fig. 3 is of the form as in equation (2)

A1 A2 A5 -> A12

3 5 20 -> 95 (2)

Where 3 is for the age range 51<=age<=75, 5 is 0<=chroniccough(weeks)<=15, 20 is for categorical, HIV=Negative and 95 is TBtype = PTB(Pulmonary Tuberculosis), it is encoded as in equation (3).

A1 A2 A3 A4 A5 A6 A7 A8.. A12

0010 0000 010 0100 0001 ..... 111 (3)

Totally twelve attributes are used for TB diagnosis. As explained in equation (1) the first two bits in equation (3) represent tag bits and last two/one bit represent the column number (value) of each range (interval) for numerical and categorical attribute respectively. In equation (3) numerical attribute A1 is antecedent(00) and its descretized range is 3 which is assigned 10 column number in our encoding. Hence its value after encoding is 0010. The same is repeated for all types of attributes.

(1) {5 11 24} -> {95} 100.0
(2) {20 24} -> {95} 100.0
(3) {5 11 20 24} -> {95} 100.0
(4) {11 94} -> {5} 99.54
(5) {11 20 22} -> {5} 99.42
(6) {11 18} -> {5} 99.38
(7) {5 20 24 95} -> {11} 97.98
(8) {24} -> {11} 97.84
(9) {24} -> {11 95} 97.84
(10) {10 11} -> {5 20} 96.79
(11) {24 95} -> {5 11} 96.77
(12) {20 95} -> {5 11} 96.73
(13) {5 11 95} -> {20 24} 95.62

Fig. 3 – Sample output from Apriori algorithm

5.3. FITNESS FUNCTION

Three sets of measures have been used as the main criteria in designing the fitness function. They are support, confidence and Lift. All the three measures have been combined along with the user defined weights assigned based on their importance in this objective function. Support is given as support(XUY)=n(XUY) / n where n is the total number of transactions in the database and n(XUY)is the number of transactions that contain both the item set X and Y. Confidence is conditional probability, for an Association rule X=>Y and defined as confidence(X=>Y) = support(XUY) / support(X). Lift assesses the degree to which the occurrence of one lifts the occurrence of the other.

lift(X=>Y) = confidence(X=>Y) / support(Y)

The proposed fitness function f(x) is specified in equation (4).

$$f(x) = (W_1*\text{Support}+W_2*\text{Confidence}+W_3*\text{Lift}) \div (W1*W2*W3) \quad (4)$$

where W_1, W_2 and W_3 are user defined weights.

6. RESULTS AND DISCUSSIONS

The experiment was conducted on the above described data source using Matlab. The population size was fixed to 30, crossover rate=0.8, mutation rate was fixed to 0.02 and number of generations=200. Compared to Fig. 3 Apriori rules which are too large, GA results in one single converged rule after running it once. Table 2 lists some of the sample subset of quality rules obtained after running Genetic Algorithm several times. Each rule in Table 2 describes the quality better than all the 227 rules obtained through Apriori. Rule one in Table 2 describes that if the TBtype is PTB (Pulmonary Tuberculosis) then the patients HIV status is Negative. Rule two and three shows the close association between cough, prolonged fever, weightloss, age and RPTB (Retroviral PTB). All the rules described in Table 2 have been verified by the doctor.

7. CONCLUSION

Association rule mining is a very useful technique in the application of medical data. It helps doctors in finding out the association of one attribute with the other. Since it generates large number of rules, it can be optimized to quality rules using Genetic Algorithm. Tuberculosis dataset with a few symptoms as attributes along with HIV as one of the attribute has been used in this experiment. We have applied a simple genetic algorithm to improve the quality of Tuberculosis Association rules. There is no restriction on the number of consequents. In the current approach datatypes such as discrete, continuous and categorical items have been handled. Binary type of encoding has been adopted. The rules which have been generated are promising, giving us better support, confidence and lift. We would like to further extend the work with other types of optimization algorithms.

8. ACKNOWLEDGEMENT

Our thanks to KIMS Hospital, Bangalore for providing the valuable real Tuberculosis data and Hospital Principal Dr. Sudharshan and his staff for giving permission to collect data from the Hospital.

Table 2. List of some of the optimized rules obtained

Selected Rules in discretized values	Corresponding attribute values
'{95} -> {24}'	TBtype=PTB -> HIV=Negative
'{20} -> {5}'	Bloodcough=null -> chroniccough(weeks)<39.0
{10 14} -> {2 6 18 22 96}	weightloss=null and intermittentfever(days)>=273.75 -> 24.25<=Age<47.5 and 39.0<=chroniccough(weeks)<78.0 and nightsweats=null and chestpain=null and TBtype=retroviralPTB
'{2 14 18 96} -> {6 10 22 23 94}'	24.25<=Age<47.5 and intermittentfever(days)>=273.75 and nightsweats=null and TBtype=retroviralPTB -> 39.0<=chroniccough(weeks)<78.0 and weightloss=null and chestpain=null and HIV=positive and wheezing=null
'{6 20 96} -> {2 14 22}'	39.0<=chroniccough(weeks)<78.0 and Bloodcough=null and TBtype=retroviralPTB -> 24.25<=Age<47.5 and intermittentfever(days)>=273.75 and chestpain=null
'{22} -> {20}'	chestpain=null -> Bloodcough=null
'{22} -> {5}'	chestpain=null -> chroniccough(weeks)<39.0

9. REFERENCES

[1] *HIV Sentinel Surveillance and HIV Estimation*, New Delhi, India, National AIDS Control Organization, Ministry of Health and Family Welfare, Government of India, 2006, http://www.nacoonline.org/Quick_Links/HIV_ Data/ Accessed 06 February, 2008.

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swamy, Mining association rules between sets of items in large databases, *Proc. ACM SIGMOD International conference on management of data*, 22 (2), 1993, pp. 207-216.

[3] Rakesh Agrawal and Ramakrishnan Srikant, Fast algorithms for mining association rules in large databases, *Proc. VLDB conference*, September 12-15, 1994, pp. 487-499.

[4] A. Savasere, E. Omiecinski, S. Navathe, An efficient algorithm for mining association rules in large database, *Proc. of the 21st VLDB Conference*, 1995, pp. 432–444.

[5] H. Toivonen, Sampling large databases for association rules, *Proc. of the 22nd VLDB Conference*, 1996, pp. 134–145.

[6] S. Birn, R. Motwani, J.D. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, *Proc. of the ACM SIGMOD*, 1997, pp. 255–264.

[7] D.I. Lin, Z.M. Kedem, Pincer search: a new algorithm for discovering the maximum frequent set, *Proc. of the 6th International Conference on Extending Database Technology: Advances in Database Technology*, 1998, pp. 105–119.

[8] D.L. Yang, C.T. Pan, Y.C. Chung, An efficient hash-based method for discovering the maximal frequent set, *Proc. of the 25th Annual International Conference on Computer Software and Applications*, 2001, pp. 516–551.

[9] R.J. Kuo, C.M. Chao, Y.T. Chiu, Application of particle swarm optimization to association rule mining, *Applied Soft Computing, Elsevier*, Vol. 11, 2011, pp. 326-336.

[10] M. Saggar, A.K. Agrawal, A. Lad, Optimization of association rule mining using improved genetic algorithms, *Proc. of the IEEE International Conference on Systems Man and Cybernetics*, Vol. 4, 2004, pp. 3725–3729.

[11] Halina Kwasnicka and Kajetan Switalski, Discovery of association rules from medical data – classical and evolutionary approaches, *Proc. of XXI Autumn Meeting of Polish Information Processing Society*, 2005, pp. 163–177.

[12] A. Ghosh, B. Nath, Multi-objective rule mining using genetic algorithms, *Information Sciences*, (163) 1-3 (2004), pp. 123-133.

[13] Sufal Das & Banani Saha, Data quality mining using genetic algorithm, *International Journal of Computer Science and Security, (IJCSS)*, (3) 2(2009), pp. 105-112.

[14] Peter P. Wakabi-Waiswa, Venansius Baryamureeba, Extraction of interesting association rules using genetic algorithms, *International Journal of Computing and ICT Research*, (2) 1 (2008), pp. 26-32.

[15] Orhan Er, Feyzullah Temurtas and A.C. Tantrikulu, Tuberculosis disease diagnosis

using artificial neural networks, *Journal of Medical Systems, Springer*, 2008, DOI 10.1007/s10916-008-9241-x online.

[16] M. Sebban, I. Mokrousov, N. Rastogi and C. Sola, A data-mining approach to spacer oligo nucleotide typing of mycobacterium tuberculosis, *Bioinformatics, Oxford University Press*, (18) 2 (2002), pp. 235-243.

[17] Rethabile Khutlang, Sriram Krishnan, Ronald Dendere, Andrew Whitelaw, Konstantinos Veropoulos, Genevieve Learmonth, and Tania S. Douglas, Classification of mycobacterium tuberculosis in images of ZN-stained sputum smears, *IEEE Transactions on Information Technology in Biomedicine*, (14) 4 (2010), pp. 949-957.

[18] G.E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison Wesley, New York, 1989.

[19] A.A. Freitas, A Survey of evolutionary algorithms for data mining and knowledge discovery, *Advances in Evolutionary Computing: Theory and Applications*, 2003, pp. 819–845.

[20] Jesus Alcala-Fdez, Rafael Alcala, Mario Jose Gacto and Francisco Herrera, Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms, *Fuzzy Sets and Systems*, (160) 7 (2009), pp. 905-921.

[21] Cristiano Pitangui, Gerson Zaverucha, Genetic based machine learning: merging Pittsburgh and Michigan, an implicit feature selection mechanism and a new crossover operator, *Proceedings of the Sixth International Conference on Hybrid Intelligent Systems* (HIS'06), 2006.

[22] Asha T., S. Natarajan, and K.N.B. Murthy, Association rule based tuberculosis disease diagnosis, *Proceedings of International Conference on Digital Image Processing (ICDIP'2010) February 26-28, 2010 SPIE*, Singapore, 7546, 75462Y1-6.

[23] Parameshvyas Laxminarayan, Sergio A. Alvarez, Carolina Ruiz, and Majaz Moonis, Mining statistically significant associations for exploratory analysis of human sleep data, *IEEE Transactions on Information Technology in Biomedicine*, (10) 3 (2006), pp. 440-450.

[24] Carlos Ordonez, Edward Omiecinski, Cesar A. Santana, et al., Mining constrained association rules to predict heart diseases, *Proc. ICDM*, November, 2001, pp. 433-440.

[25] Carlos Ordonez, Cesar A. Santana, Levien de Braal, Discovering interesting association rules in medical data, *Proc. ACM DMKD*, 2000, pp. 78-85.

[26] B. Liu, W. Hsu, S. Chen, Y. Ma, Analyzing the Subjective Interestingness of Association Rules, *IEEE Intelligent Systems, 2000.*

[27] M. Pei, E.D. Goodman, W.F. Punch, Feature extraction using genetic algorithm, *Proceedings of International Symposium on Intelligent Data Engineering and Learning (IDEAL'98)*, 1997.

[28] Sufal Das, Bhabesh Nath, Dimensionality reduction using association rule mining, *IEEE Region 10 Colloquium and Third International Conference on Industrial and Information Systems (ICIIS 2008)*, December 8-10, 2008, IIT Kharagpur, India.

[29] M. Anandhavalli, Suraj Kumar Sudhanshu, Ayush Kumar and M.K. Ghose, Optimized association rule mining using genetic algorithm, *Advances in Information Mining*, (1) 2
(2009), pp. 01-04.

[30] David Beasley et al., An overview of genetic algorithms, Part 1 & 2, University Computing, (15) 2 & 4 (1993), pp. 58-69 & pp. 170-181.

[31] Kazuo Sugihara, Measures for performance evaluation of genetic algorithms, *Proceedings of 3rd Joint Conference on Information Sciences (JCIS'97)*, 1997, pp. 172-175.

[32] S.M Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil, Genetic mining: using genetic algorithm for topic based on concept distribution, *Transactions on Engineering Computing and Technology*, Vol. 13, World Enformatika Society, May 2006, pp. 44-147.

[33] Mehmet Kaya, Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules, *Soft Computing*, 2006, pp. 578-586.

[34] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, II edition, Morgan Kaufmann Publishers, San Francisco, 2006.

[35] Ian H. Witten and Eibe Frank, *Data Mining Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2001.

[36] A.K. Pujari, *Data Mining Techniques*, Universities Press, 2001.

[37] Carlos Ordonez, Association rule discovery with the train and test approach for heart disease prediction, *IEEE Transactions on Information Technology in Biomedicine*, (10) 2 (2006), pp. 334-343.

[38] T.J. Chen, L.F. Chou and S.J. Hwang, Application of a data mining technique to analyze coprescription patterns for antacids in Taiwan, *Clin. Ther.*, (25) 9 (2003), pp. 2453–2463.

[39] Xiaowei Yan, Chengqi Zhang and Shichao Zhang, Genetic algorithm-based strategy for

identifying association rules without specifying actual minimum support, *Expert Systems with Applications*, Vol. 36, 2009, pp. 3066-3076.

[40] Xiaowei Yan, Chengqi Zhang, and Shichao Zhang, ARMGA: identifying interesting association rules with genetic algorithms, *Taylor & Francis*, Vol. 19, 2005, pp. 677-689.

[41] A.J. Christian, G.P. Martin, Optimization of association rules with genetic algorithm, *Proceedings of XXIX IEEE International Conference of Chilean Computer Science Society (SCCC 2010)*, USA.

[42] Ashish Ghosh, S. Dehuri and S. Ghosh, *Multiobjective Evolutionary Algorithms for Knowledge Discovery from Databases*, Book series, "Studies in Computational Intelligence", Vol. 98, Springer, 2008.

Mrs. Asha.T *obtained her Bachelors and Masters in Engg., from Bangalore University, Karnataka, India. She has her Ph.D from Visveswaraya Technological University under the guidance of Dr. S. Natarajan and Dr. K.N.B. Murthy. She has over 18 years of teaching experience and currently working as Professor in the Dept. of Computer Science & Engg., B.I.T. Karnataka, India. Her Research interests are in Data Mining, Medical Applications, Pattern Recognition, and Artificial Intelligence.*

Dr. S.Natarajan *holds Ph. D (Remote Sensing) from JNTU Hyderabad India. His experience spans 33 years in R&D and 10 years in Teaching. He worked in Defence Research and Development Laboratory (DRDL), Hyderabad, India for five years and later worked for Twenty Eight years in National Remote Sensing Agency, Hyderabad, India. He has over 50 publications in peer reviewed Conferences and Journals His areas of interest are Soft Computing, Data Mining and Geographical Information System.*

Dr. K. N. B. Murthy *holds Bachelors in Engineering from University of Mysore, Masters from IISc, Bangalore and Ph.D. from IIT, Chennai India. He has over 30 years of experience in Teaching, Training, Industry, Administration, and Research. He has authored over 60 papers in national, international journals and conferences, peer reviewer to journal and conference papers of national & international repute and has authored book. He is the member of several academic committees Executive Council, Academic Senate, University Publication Committee, BOE & BOS, Local Inquiry Committee of VTU, Governing Body Member of BITES, Founding Member of Creativity and Innovation Platform of Karnataka. Currently he is the Principal & Director of P.E.S. Institute of Technology, Bangalore India. His research interest includes Parallel Computing, Computer Networks and Artificial Intelligence.*