INTERNATIONAL JOURNAL OF COMPUTER SCIENCE ENGINEERING RESEARCH AND DEVELOPMENT (IJCSERD)
ISSN 2248 – 9363(Print)
ISSN 2248 – 9371(Online),
Volume 3, Number 2, April- May (2013), pp: 30-45

IJCSERD

© PRJ PUBLICATION

A GENETIC BASED RESEARCH FRAMEWORK TO DISCOVER OPTIMAL FREQUENT PATTERNS USING ASSOCIATION RULE MINING

Prof. V. V. R. Maheswara Rao[1], Scientist Mentor
N. Silpa[2]**,** Principal Investigator

[1, 2]Shri Vishnu Engineering College for Women, Bhimavaram, AP, India

ABSTRACT

The rapid advances in data generation, availability of automated tools in data collection and continued decline in data storage cost enabled with high volumes of data. In addition, the data is non scalable, high dimensional, heterogeneous and complex in its nature. This situation creates inevitably increasing challenges in extracting desired information. Thus, Data mining evolves into a fertile area and got the focus by many researchers and business analysts. Data mining is a methodology the blends traditional techniques with sophisticated algorithms. Among all, the association rule mining is efficient pattern discovery technique, which finds hidden, valid, novel, useful, understandable, interesting and ultimately correlated patterns in large databases. Such correlated rules create great business value to any organization as they make use in decision making process. However, in real time applications the correlation changes continuously as the source data updates dynamically. This motivation necessitates finding and updating the frequent item sets with different supports efficiently and optimally.

In order to overcome the challenges inherited in conventional association rule mining, the authors in the present paper propose an Optimal Frequent Patterns System (OFPS). The OFPS takes radically a different approach and design as a three-fold system that discovers optimal frequent patterns efficiently, using the genetic algorithm. Initially, the first-fold of OFPS focuses on preparation of domain specific data that includes data selection, cleaning, integration and transformation under the guidance of knowledge expert. Subsequently, the second-fold of OFPS emphasizes on construction of a Frequent Pattern Tree (FP-Tree) and then discovery of frequent patterns by exploring the tree in the bottom-up fashion to facilitate rapid access of individual frequent patterns quickly. The third-fold of OFPS finally concentrates on generation of optimal frequent patterns using genetic algorithm that simulates biological evaluation procedure having the self learning capability. To validate the

performance of proposed OFPS in several orders of magnitude, many experiments were conducted and results have proven this as claimed.

Keywords-Data Mining; Frequent patterns; Association rule mining; Optimizaation techniques; Genetic algorithm.

1. INTRODUCTION

The past decade has seen an explosive growth in database technology and enormous proliferation of data in every area of human endeavor. The advances in data collection, use of bar codes, RFID in commercial outlets, and automation of business transactions have flooded with lots of data. These causes have created a great demand towards data mining research to model potential and optimal systems for turning data into useful and task oriented knowledge. Extracting the knowledge from such a complex and huge amount of data efficiently and effectively is becoming a tedious process.

Knowledge Discovery in Databases (KDD) is a process of extracting valuable, unknown, valid and actionable information from large databases to make crucial business decisions. The iterative process of KDD includes Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Representation. The steps of KDD process is as shown in figure 1.

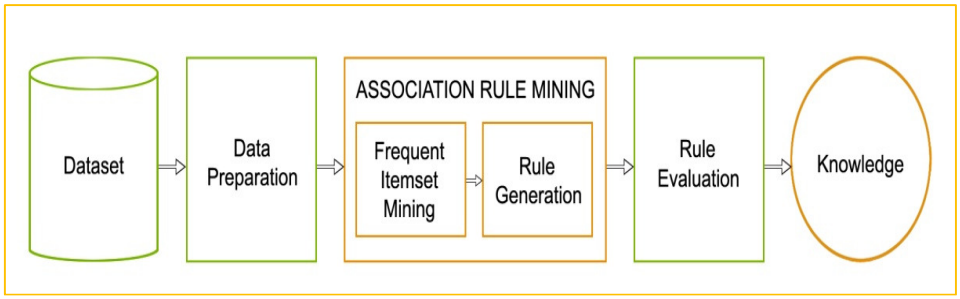

Figure 1 Steps of KDD process

Data mining is a sifting process to extract useful patterns from large amount of data, either directly in the form of knowledge that characterizes the relation between the variables of interest, or indirectly as functions that represent patterns. There have been many different techniques used to perform data mining task. Basically, these techniques are categorized into parametric techniques which follow model-based approach and non-parametric techniques which fallow data driven approach.

Specifically, non-parametric techniques are more appropriate for real world data mining applications with large amount of dynamically growing data. In addition, the recent non-parametric techniques have employed the machine learning techniques like Neural Networks, Decision Trees and Genetic Algorithms to learn dynamically. The well-known data mining functionalities like Association rule mining, Classification and Clustering are fall into non parametric techniques.

Among all functionalities of data mining, the problem of deriving association from data received a great deal of attention. Association rule mining techniques can be used to extract interesting correlations, frequent patterns and associations among a set of items in the transactional databases. The simple usage of association rule mining techniques in KDD process is discouraged as they generate more number of patterns.

Besides, the high volumes of transactional data and impractical environment of data set, the existing association mining techniques find difficulty in handling the newly emerged problems for further stages of data mining especially, the pattern analysis. To proceed towards intelligence, reducing the need of human intervention, it is necessary to integrate and entrench artificial intelligence into data mining techniques. To achieve the intelligence, soft computing methodologies seem to be a good candidate.

The soft computing models are characterized by its ability for granular computation in avoiding the concept of approximation. Basically, soft computing models provide the foundation for computational intelligence systems and further outline the basis of future generation computing systems. These models are close resemblance to human like decision making and used for modeling highly non linear data, where the pattern discovery, rule generation and learnability are typical. The Fuzzy Logic, Artificial Neural Networks, Genetic Algorithms and various combinations of these techniques have made the Soft computing paradigm. Among which, Genetic Algorithms, a biologically inspired technology and is more suitable for association mining.

This situation promotes the necessity of applying the optimization techniques to get the optimal frequent patterns has become a main motivation for the proposed work. Many authors in the literature survey introduced several soft computing methodologies and deliberately express the relevance of Genetic Algorithms in the Association Rule Mining.

Genetic algorithms are designed based on biologically inspired technology with granular computing nature and is more suitable for predictive data mining techniques. The GA is more adequate since the implicit parallelism of GA can mine the large data depositories with less time in yielding the exact optimal solution. Thus, in order to find the optimal frequent patterns the proposed work presents a genetic based research frame work using association rule mining.

The remaining paper is organized as follows. Section 2, provides a detailed review on association rule mining and the usage of genetic algorithm suitable to proposed work. The next section 3, presents the proposed work in detail. The subsequent section 4 showcases the experimental analysis of the proposed work. Finally in section 5 conclusions are mentioned.

2. RELATED WORK

The research work in this paper conducted the literature survey from 2006 to the current year with a focus on each phase of proposed system.

In 2006, Nan Jiang and Le Gruenwald [18] provided a detailed survey on association rule mining and specifically concentrated on data streams. They expressed that the conventional algorithms are inefficient with huge amount and changing distribution of data. Finally, they concluded that it is necessary to design the more efficient and user friendly mining techniques to address all performance issues in association mining. In the same year, S. Y. Wang, K. Tai, and M. Y. Wang [19] presented a versatile, robust and enhanced genetic algorithm for structural topology optimization using problem specific knowledge. In their

implementation process specifically pronounced the importance of choosing appropriate representation techniques, genetic operators and evaluation methods.

In the subsequent year 2007, Rong Gang, Liu Jin-feng, Gu Hai-jie [16] put forward a new concept, dynamic association rule which can describe the regularities of change-over-time in the association rules. It contains not only a support and confidence but also a support vector and confidence vector. In the same year, Ansaf Salleb-Aouissi, Christel Vrain and Cyril Nortet [17] proposed a mining quatitative association rules system that dynamically discovers good intervals in association rules by optimizing both support and confidence based on genetic algorithm. Their results are evident that genetic algorithm is suitable in handling optimization problem of association rule mining.

All the range in 2008, J L Balcazar [14] studied and explored the concept of redundancy among association rules from a fundamental perspective. They discussed several existing alternative definitions of redundancy between association rules and provided new characterizations and relationships among them. They also provided a sound and complete calculus to construct deduction scheme for redundancy rules. During this year, S. Ventura, C. Romero, A. Zafra, J. A. Delgado, and C. Hervas [15] designed a framework that can apply to maximize reusability and availability of evolutionary computation with a minimum effort in web mining. The heavily demanding computational performance is an open problem as earmarked in their future research work.

Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K. in 2009 [12] explained the importance of negative association rules in the association rule mining. They provided a general overview on genetic algorithm and its relevance to get optimized association rules. Hyunchul Ahn, Kyoung-jae Kim [13] reviewed prior studies on optimization techniques for several systems. They further examined genetic approach for optimization of feature weights and relevant instances for similarity calculations. They also mentioned in their limitations that the size of the population and the number of generations for genetic algorithm is very huge. Thus, reducing the size of population and number of generations for genetic algorithm is an open challenge.

In the year 2010, Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar and Partha Pratim Sarkar [9] reviewed the prior studies of association rule mining and demonstrated the usage of genetic algorithms in finding the frequent item sets based on appriori algorithm. They deliberatively expressed that the incorporation FP Tree with the genetic algorithm is a future research path. Again in 2010, Mehmet Kaya [10] proposed a novel method using multi objective evolutionary algorithm that extracts the patterns automatically. This method applied on dataset with a sequential character. Their experiments demonstrated on real datasets which exhibit good performance in terms of accuracy. The methodology of automatic extraction is a promising future research as mark down in their conclusions.

In 2011, Diana Martın, Alejandro Rosete, Jesus Alcala-Fdez and Francisco Herrera [7] extended the well-known multi-objective evolutionary algorithms to perform learning of the intervals of attributes and a condition selection in order to mine a set of optimum association rules with accuracy.

During 2012, Xiaoyan Sun, Lei Yang, Dunwei Gong and Ming Li, [4] studied that collective intelligence extracted from multiple users enhance the performance of GA. The performance of the proposed algorithm is emperically validated on its application to fashion design system. They felt that designing evolutionary algorithm is a promising research direction in the knowledge discovery process as mentioned in their future research directions.

Recently in 2013, Johannes K. Chiang, Rui-Han Yang [1] proposed an approach which includes a novel data structure and an efficient algorithm for mining association rules on various granularities. However, their test results shown its performance, efficiency and scalability better than the current approaches. But the effects of perceived issues and potential development of data mining and concept description are worthy of further investigation. In current year, 2013, Gaurav Dubey, Arvind Jaiswa [2] have dealt the challenge of association rule mining problem in finding frequent itemsets using GA based method. However, they noticed that a more extensive empirical evaluation of their proposed method is a promising future research.

Many of the earlier authors as observed in the literature have explained the importance and efficiency of genetic based approach in the process of discovering optimal frequent patterns, which has been considered as the formal basis for the present work that motivate the authors to define the proposed system. To develop more efficient and optimal techniques to serve the increasing demands of each organization has become the prime motivation to the present work.

3. PROPOSED OPTIMAL FREQUENT PATTERNS SYSTEM (OFPS)

In order to overcome the challenges inherited in earlier works, the authors propose an Optimal Frequent Patterns System (OFPS) that takes radically a different approach and designed as a three-fold system. Initially, the first-fold of OFPS focuses on preparation of data that includes data selection, cleaning, integration and transformation under the guidance of a knowledge expert. Subsequently, the second-fold of OFPS emphasizes on construction of a Frequent Pattern Growth Tree, and then discovers the frequent patterns by exploring the tree in the bottom up fashion to facilitate rapid access of individual frequent patterns quickly. The third-fold of OFPS finally concentrates on generation of optimal frequent patterns using Genetic Algorithm. The architecture of OFPS is as shown in figure 2.

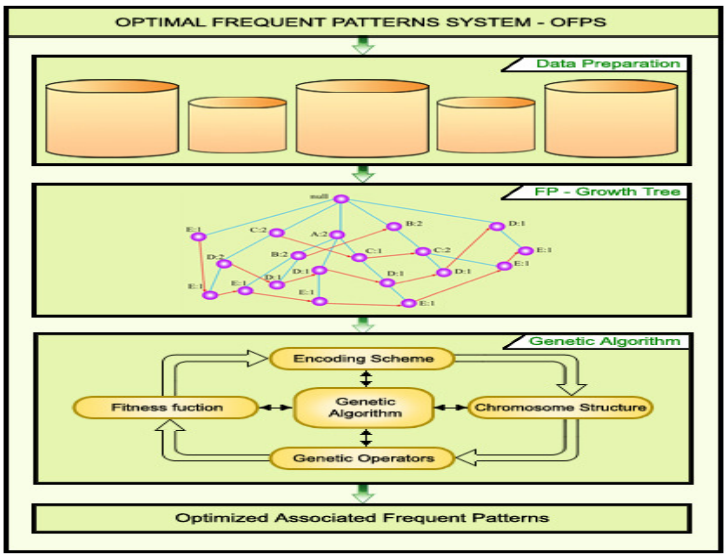

Figure 2. Architecture of Optimal Frequent Patterns System

3.1 DATA PREPARATION

The first fold of OFPS focuses on preparation of desired data under the guidance of a knowledge expert as raw data, highly predisposed to noise, missing values and inconsistency. This data preparation stage is the most important phase in the KDD process and is critical in successful extraction of desired data. The data preparation helps to improve the efficiency and ease of any data mining technique. The task of data preparation consumes a bulk amount of effort in the entire data mining investigation. The data preparation of OFPS covers all the activities including data collection, data cleaning, data integration, data transformation and data reduction to construct the final dataset from the initial raw data.

Data Collection: The objective of data mining technique is the only key driven force for collecting the data. The data collection task is performed on the basis of input attributes drawn from the desired task. This activity includes significance to the data mining goals, quality and technical limitations such as limits on data volume or data types. It is very important, however, to understand how data collection affects the data mining techniques, since such a prior knowledge is also useful for the final interpretation of results.

Data Cleaning: Data cleaning, also called data cleansing, deals with detecting and removing the incomplete, noisy and inconsistent data in order to improve the quality of data. This data preparation activity of OFPS is particularly required when integrating large and real-world heterogeneous databases. Initially, to fill the missing values, OFPS employs a popular strategy "Use the most probable value to fill in the missing value". This strategy uses most of the information from the present data to predict missing values that are determined with Bayesian approach. Subsequently, OFPS designates a linear regression model to smoothing the noisy data. The mathematical equation derived using linear regression model fit the data and helps to smooth out the noise. Finally, OFPS adopts the concept of functional dependencies between attributes to resolve the inconsistencies.

Data Integration: Data integration is a process that combines data from multiple distributed sources into a coherent data store. It specifically, aims at increasing the completeness, conciseness and correctness of the data which is fed to the data mining techniques. The completeness measure concentrates on the number of attributes while the conciseness identifies the uniqueness of attribute in the integrated data. Additionally, correctness measure focuses on confirmation of integrated data to the real world. The data integration activity of OFPS primarily resolves heterogeneity and schema level by establishing semantic mapping among contents of multiple data sources. The next level it resolves heterogeneity at instance level by identifying the records that refer to the real world entity.

Data Transformation: The data transformation consolidates the data into a single desired form which is readily fed to the mining technique. The data transformation activity of OFPS involves normalization, aggregation and generalization of data. The normalization concentrates to scale the data in a small specified range. The aggregation performs functions that are applied to the data for summarization. In generalization of data the raw data is replaced by higher level concepts using concept hierarchy.

Data Reduction: The data reduction is a technique that reduces volume of data set much smaller, at closely maintains the integrity of original data. The data reduction activities of OFPS uses attribute subset selection strategy for data reduction. This strategy reduces the data set size by removing irrelevant attributes. The role of this strategy is to find a minimum set of attributes such that, the resulting probability distribution is as close as possible to the original distribution obtained using all attributes.

The whole activity of data preparation elevates the quality of the data set to the required level by the data mining techniques. This resultant processed data will be fed to the next fold of OFPS finding optimal frequent patterns efficiently.

3.2 FPGROWTH TREE

The second-fold of OFPS emphasizes on finding complete set of frequent patterns without candidate generation by employing FP-growth algorithm, thus improving performance. It is one of the fastest and most popular algorithms of current age and adopts divide-and-conquer strategy. It is based on a compact prefix tree representation called a Frequent Pattern Tree (FP-Tree), which retains patterns association information. The construction of a FP-Tree is a compressed representation of complete data by reading one pattern at a time and mapping each pattern on to a path at a single scan. The discovery of frequent patterns is by exploring the FP-Tree using the pointers which connects between the nodes that have same patterns in bottom up fashion and that helps to facilitate rapid access of individual frequent patterns in the tree.

Initially, with the first scan of the transactional database, the FP-growth algorithm determines frequencies of each item and eliminates that are not frequent individually with the user specified minimum support. In addition, the items in each patterns are sorted in descending order with respective their frequencies. Although, the algorithm does not depend on specific order, the experimental result showed by [19] indicates that the execution time with descending order is shorter than random order.

- Scan data and find support for each item
- Discard infrequent items
- Sort frequent items in decreasing order based on their support

Later, the FP-growth algorithm makes a second scan of the data to construct the FP-Tree. After reading the initial pattern, the nodes of the tree are labeled accordingly and a path is formed to encode the pattern. After reading the new pattern and no common prefix is found, a new set of nodes is created, labeled and a path is formed by connecting all the nodes in the pattern. While reading new pattern and common prefix is found, the frequency count for the node is incremented and an overlap path is formed. This process continues until every transaction has been mapped on to one of the paths given in the FP-Tree.

Algorithm 1: FP-Tree construction:

1. Scan the transaction database once. Collect F, the set of frequent patterns, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequent patterns.

2. Create the root of an FP-tree, T, and label it as "null". For each transaction Trans in database do the following:
   - Select the frequent patterns in Trans and sort them according to the order of FList. Let the sorted frequent-pattern list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree([p | P], T).
   - The function insert tree([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name, then increment N 's count by 1; else create a new node N , with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree(P, N) recursively

Finally, FP-growth algorithm, concentrates on generating frequent patterns from FP-Tree by exploring the tree in the bottom-up fashion. This strategy finds the frequent patterns ending with a particular item, by examining only the path ending with the same item. This process continues until all the paths associated with all nodes are processed. These paths are accessed rapidly since FP-Tree stores the associated item information.

Example for FP-Tree:

An example of processed transactional data set as shown in table 1 is taken as input to demonstrate the construction of FP-Tree. Item wise frequency count is calculated and shown in table 2. On reading the sessions one by one FP-Tree is constructed and complete tree is shown in figure 3.

Table 1 Snapshot of Sessions

Transaction Id	List of Item Ids
1	I_1, I_2
2	I_2, I_3, I_4
3	I_1, I_3, I_4, I_5
4	I_1, I_4, I_5
5	I_1, I_2, I_3
6	I_1, I_2, I_3, I_4
7	I_1
8	I_1, I_2, I_3
9	I_1, I_2, I_4
10	I_2, I_3, I_5

Table 2 Item Wise Frequency Count

Item Id	Frequency Count
I_1	8
I_2	7
I_3	6
I_4	5
I_5	3

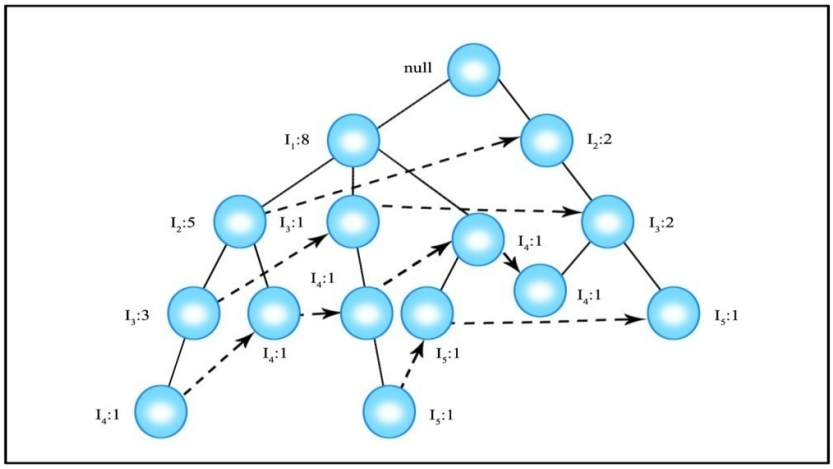

Figure 3 Complete FP-Tree on reading all Transactions

3.3 OFPS-GENETIC APPROACH

The third-fold of OFPS finally pays attention to generate optimal frequent patterns using OFPS-Genetic Algorithm that simulates biological evaluation procedure having the self learning capability. It explicitly strives to evolve concise patterns that can be directly inspected and interpreted. In addition, the OFPS is an advanced optimization technique outperforms the conventional association mining algorithms by several orders of magnitude. The OFPS considers each stage of genetic algorithm in view of association rule mining. The stage-by-stage process of genetic algorithm is shown in the third part of figure 2.

The encoding strategy is an initial and the toughest stage of genetic algorithm that finds the initial population from frequent patterns generated by FP-Growth algorithm to initiate the process. Then fitness function evaluates the survival frequent patterns by the theory of evolution from the initial population and generates the next biological population. In the next stage the biologically inspired genetic operators create a new and potentially better population. Finally, the end function of genetic algorithm terminates the process as and when an acceptable set of optimal frequent patterns is found or after the lapse of a fixed time interval.

OFPS-Encoding Scheme: The encoding scheme is a process of representing output generated by FP-Growth algorithm into a suitable form to the genetic algorithm. It is an important issue in genetic process as it plays a critical role to arrive at best performance of

algorithm as robust as possible. GA uses various encoding schemes like tree encoding, permutation encoding, binary encoding etc., here OFPS adopts binary encoding.

Consider following example of pattern $\{I_1, I_4, I_5\}$ is encoded as a binary chromosome of length 5 and is shown in Figure 4. The presence of an item in a pattern is coded as 1, otherwise as 0.

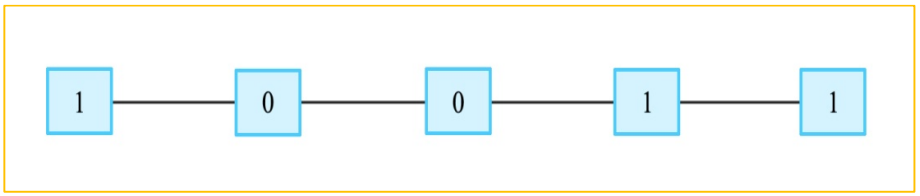

Figure 4. Example Binary Chromosome

OFPS-Fitness Function: The fitness function evaluates the optimality of a pattern so that a particular pattern is ranked against all other patterns. It is an essential step in the overall process of genetic approach as it plays a key role to assess the survival capacity of a pattern. The OFPS employs a robust fitness function which is designed based on confidence factor (CF) and completeness measure (CM). These measures are calculated using values of contingency table for given pattern.

Consider an example of 2 x 2 contingency table of a given pattern, to calculate both CF and CM for an associated pattern $\{I_1, I_2\}$ is as shown in table 3.

Table 3 2 X 2 Contingency Table of a Pattern

	$\boldsymbol{I2}$	$\overline{\boldsymbol{I2}}$
$\boldsymbol{I1}$	f_{11}	f_{10}
$\overline{\boldsymbol{I1}}$	f_{01}	f_{00}

Where,

- I_1, I_2 denote the items in an associated pattern $\{I_1, I_2\}$
- f_{11} denotes the number of associated patterns satisfying both I_1 and I_2
- f_{10} denoted the number of associated patterns satisfying I_1 but not I_2
- f_{01} denotes the number of associated patterns satisfying I_2 but not I_1
- f_{00} denoted the number of associated patterns not satisfying both I_1 and I_2

Confidence Factor, **CF** = $\{f_{11} / (f_{11} + f_{01})\}$ Mod1
Complete measure, **CM** = $\{f_{11}/ (f_{11}+ f_{10})\}$ Mod1
Thus, **Fitness function** = (CF *CM) Mod1

In this fitness function, Mod operation with 1 assures the range of fitness function value, which is [0...1]. The value of fitness function represents the accuracy rate of frequent pattern optimality. The fitness function is computed after each step until the genetic algorithm is terminated.

OFPS-Operators: The biologically inspired genetic operators of OFPS are applied on initial population of frequent patterns as chromosomes to generate possible better new offspring. The Selection, Crossover and Mutation are set of operators designated by OFPS which transforms individual chromosomes stochastically. Each chromosome has an associated value called fitness function that contributes in the generation of new population by genetic operators. At each generation, the OFPS utilizes the fitness function values to evaluate survival capacity of each chromosome. The OFPS operators create a new set of population iteratively to improve on the current fitness function values by using old ones.

Selection: The selection operator decides the number of times a particular individual chromosome is chosen for reproduction from current population as a mating pool for further OFPS operations. The number of individual chromosomes obtain for the next generation is directly proportional to its fitness value, there by mimic the natural selection procedure. This scheme is commonly called the proportional selection scheme. Roulette wheel parent selection, stochastic universal selection and binary tournament selection are some of the most frequently used selection procedures.

Here the OFPS deploys the roulette wheel parent selection procedure. This wheel as many slots as population size where the size of the slot is proportional to the relative fitness of corresponding frequent pattern chromosome in the initial population as demonstrated in figure 5. An individual frequent pattern is selected by spinning the roulette and noting the position of the marker when the roulette stops. Thus, the number of times the selection of individual frequent pattern is proportional to its fitness function value in the population.

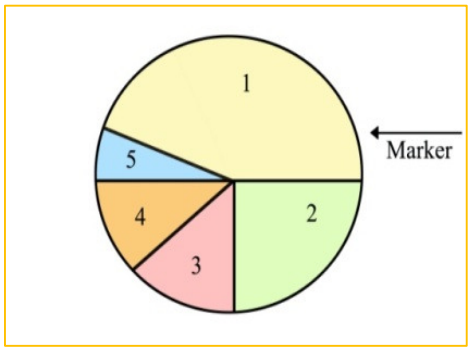

Figure 5 Example of Roulette Wheel Parent Selection

Crossover: The main purpose of the crossover is to exchange information between randomly selected parent chromosomes by recombining parts of their genetic materials. This operation performed probabilistically, combines best characteristics of parents to produce offspring for the next generation. Single-point crossover, two-point crossover, multiple-point crossover, shuffle exchange crossover and uniform crossover are the most frequently used crossover techniques.

The OFPS designates single-point crossover technique. Here, the members of selected frequent patterns in the mating pool are first paired at random then, for performing crossover on a pair, an integer position K known as crossover point is selected randomly between one end S-1 where S is the size of the frequent pattern. Two new patterns are created by swapping

all characters from the position K+1 to S. For example, the two parent patterns depicted with two different colors and crossover point are shown in figure 6.

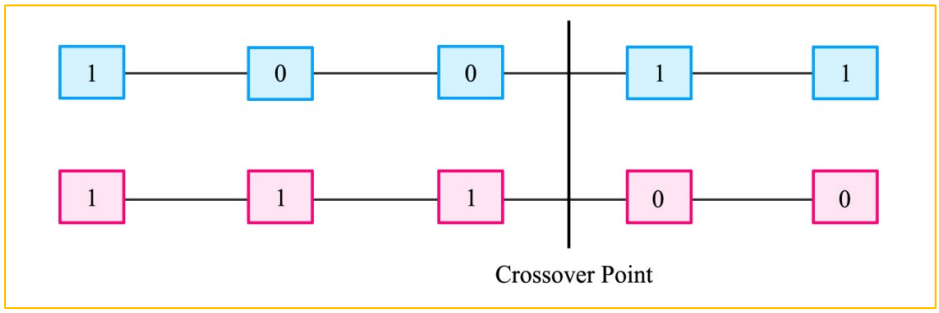

Figure 6 Single Point Crossover Operation before crossover

Finally, it performs crossover operation on a pair of patterns at the crossover point. Then, the parts of two parent patterns after the crossover point are exchanged to form new offspring as shown in figure 7.

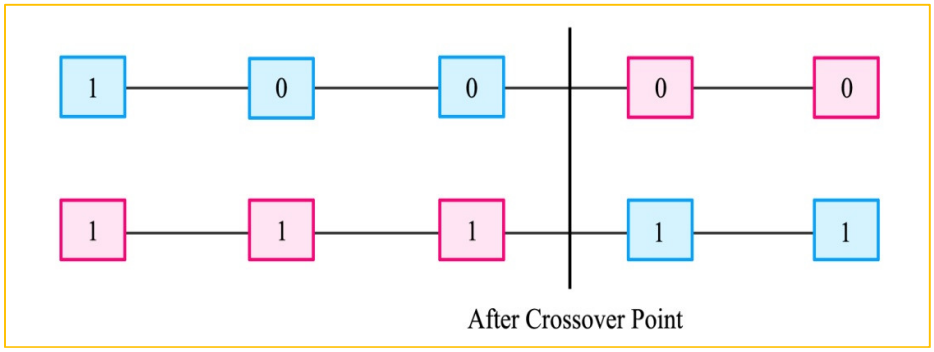

Figure 7 Single Point Crossover Operation after crossover

Mutation: Mutation is the process by which a random alteration in the genetic structure of a chromosome takes place. The main aim of mutation is to introduce genetic diversity into the new population. In some problems, it may so happen that the optimum solution resides other than initial population. In such problems only mutation can possibly direct towards optimal solutions. Mutating a binary gene defined in a variety of ways in the literature.

Here OFPS uses binary bit-by-bit mutation. An example of binary bit-by-bit mutation is shown in figure 8. Here, the positions 2 and 4 of the chromosome pattern have been subjected to mutation.

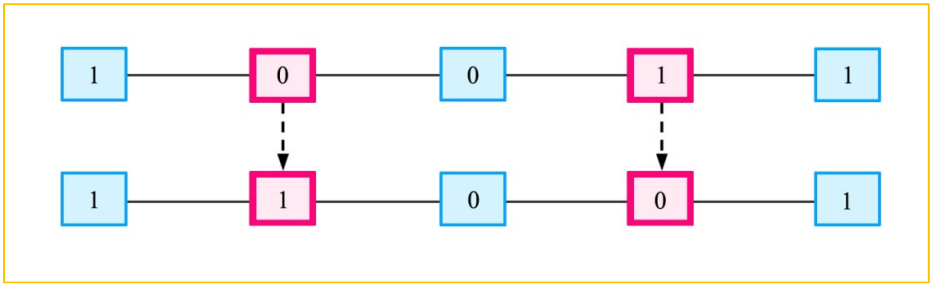

Figure 8 Process of bit-by-bit mutation

OFPS-Genetic Algorithm:

Step 01. Start
Step 02. Load a sample of records from the database that fits into memory
Step 03. Apply FP-Growth algorithm to find the frequent patterns with the minimum support. Suppose S is set of the frequent patterns set generated by FP-Growth algorithm.
Step 04. Set Q=Ø where Q is the output set, which contains the all Frequent patterns
Step 05. Set the Input termination condition of Genetic Algorithm
Step 06. Represent each frequent patterns of S as binary encoding
Step 07. Select the two members (string) from the frequent pattern
Step 08. Apply GA operators, crossover and mutation on the selected members (string) to generate the Optimal Frequent patterns
Step 09. Find the fitness function value
Step 10. If (fitness function value > min confidence) then
Step 11. Set Q = Q U {x =>y}
Step 12. If the desired number of generations is not completed, then go to Step 3.
Step 13. Stop

4. EXPERIMENTAL ANALYSIS

The proposed OFPS is experimented on several synthesized data sets under standard execution environment. For the OFPS-Genetic Algorithm, the frequent patterns generated by FP-Growth are given as input to start the process.

A) The OFPS compared with the Apriori and FP-Growth in terms of execution performance. The experimental results indicate that noticeable improvement of OFPS performance over the Apriori and FP-Growth techniques as shown in figure 9.

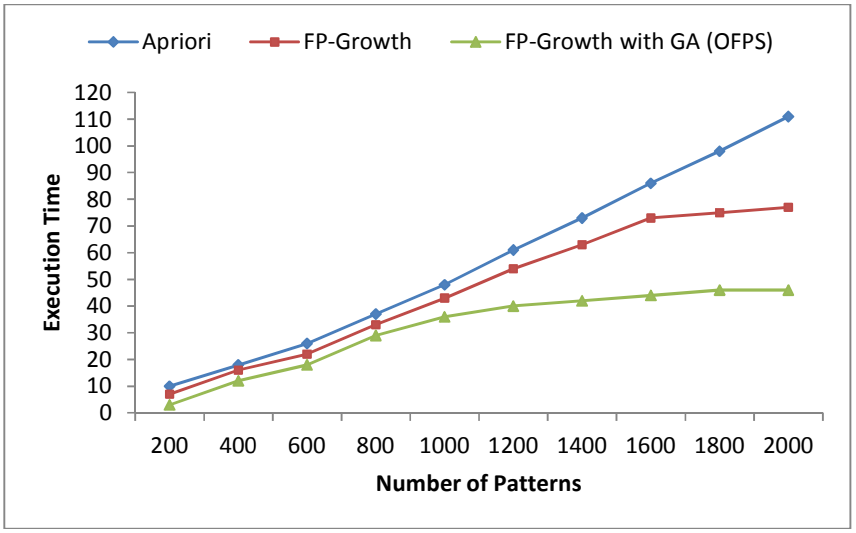

Figure 9 Efficiency Comparison of OFPS with earlier Techniques

B) The optimal patterns generated by OFPS compared with the frequent patterns generated by FP-Growth Algorithm, and the graph is depicted as shown in figure 10. The results indicate that noticeable invalid frequent patterns are identified and correspondingly the reduced number of sustainable optimal patterns is shown in the figure for each data set. The results evidently infer that the proposed OFPS has relevance and promising future to arrive at optimal solution intelligently in the association mining.

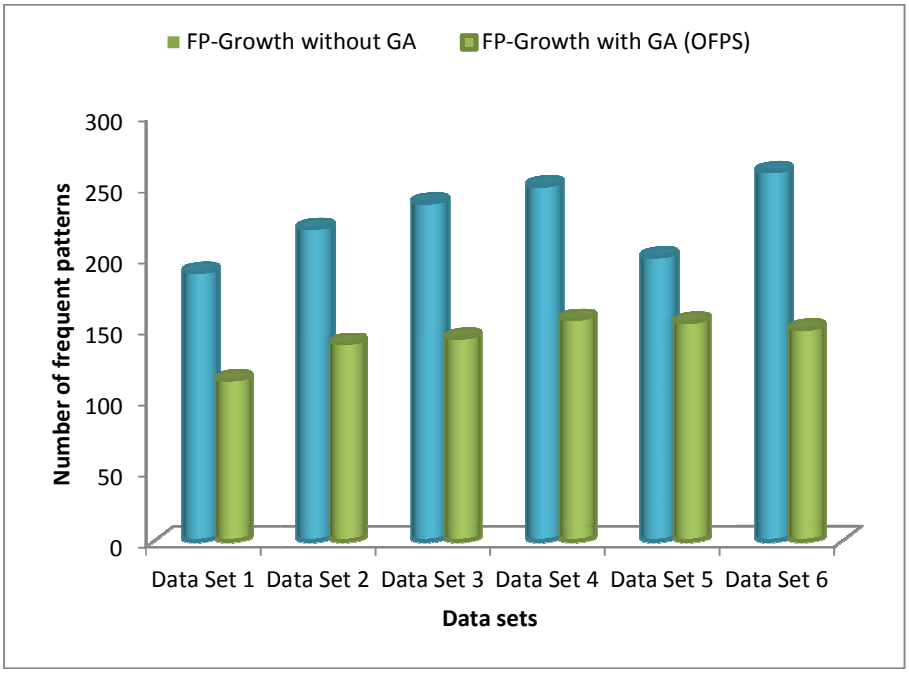

Figure 10 Performance Comparison of OFPS over FP-Growth Algorithm

5. CONCLUSIONS

The present model has proven the relevance of genetic algorithm in the identification of the optimal frequent patterns. The results are evident that the proposed OFPS has a promising future to arrive at optimal solution in the association rule mining. The binary encoding strategy of the proposed system exactly represents the each frequent pattern generated by FP-Growth algorithm as chromosome and in turn rightly prepares the initial population. The confidence factor and completeness measure of fitness function evaluates the survival of new population beyond the support and confidence frame work, yields high accuracy of optimality. The nature of biological diversity of OFPS prevents the population from stagnating at any local solution. Moreover, the stochastic process of OFPS, assures the optimal solution always.

ACKNOWLEDGEMENTS: The authors would like to thank the Department of Science & Technology (DST), Ministry of Science & Technology, Government of India under Women Scientist Scheme A (WOS-A) for providing the fund to this research. The authors also recorded their acknowledgements to the authorities of Shri Vishnu Engineering College for Women, Bhimavaram, A.P., India for their constant support and cooperation.

6. REFERENCES

[1]. Johannes K. Chiang, Rui-Han Yang, "Multidimensional Data Mining for Discover Association Rules in Various Granularities", IEEE Conference Publications, pp: 1-6, 2013.

[2]. Gaurav Dubey, Arvind Jaiswal, "Identifying Best Association Rules and Their Optimization Using Genetic Algorithm", International Journal of Emerging Science and Engineering (IJESE), Volume-1, Issue-7, pp: 91-96, 2013.

[3]. V.V.R. Maheswara Rao and Dr. V. Valli Kumari "An Intelligent Optimal Genetic Model to Investigate the User Usage Behaviour on World Wide Web", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, pp: 33-48, 2013.

[4]. Xiaoyan Sun, Lei Yang, Dunwei Gong and Ming Li, "Interactive Genetic Algorithm Assisted with Collective Intelligence from Group Decision Making", IEEE World Congress on Computational Intelligence, pp: 1-8, 2012.

[5]. Sanat Jain, Swati Kabra " Mining & Optimization of Association Rules Using Effective Algorithm", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 4, pp: 281-285, 2012.

[6]. K. Poornamala and R. Lawrance "A General Survey on Frequent Pattern Mining Using Genetic Algorithm", Journal on Soft Computing, Volume 03, Issue 01, 2012.

[7]. Diana Martın, Alejandro Rosete, Jesus Alcala-Fdez and Francisco Herrera, "A Multi-Objective Evolutionary Algorithm for Mining Quantitative Association Rules", IEEE Conference Publications, pp: 1397-1402, 2011.

[8]. Rakhi Garg, P.K. Mishra "Exploiting Parallelism in Association Rule Mining Algorithms" International Journal of Advancements in Technology http://ijict.org/ ISSN 0976-4860, Vol 2, No 2, 2011.

[9]. Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, 2010.

[10]. Mehmet Kaya, "Automated extraction of extended structured motifs using multi-objective genetic algorithm" Expert Systems with Applications, Volume 37, Issue 3, pp: 2421-2426, 2010.

[11]. V.V.R. Maheswara Rao, Dr. V. Valli Kumari and Dr. K.V.S.V.N. Raju "A Plausible Comprehensive Web Intelligent System for Investigation of Web User Behaviour Adaptable To Incremental Mining" International Journal of Database Management Systems (IJDMS) Vol.2, No.3, 2010.

[12]. Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K. "Optimized association rule mining using genetic algorithm", Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, pp-01-04, 2009.

[13]. Hyunchul Ahn, Kyoung-jae Kim, "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, Applied Soft Computing, Volume 9, Issue 2, pp: 599–607, 2009.

[14]. J L Balcazar, "Redundancy, Deduction Schemes, and Minimum-Size Bases for Association Rules" Pascal Report 4259, 2008.

[15]. S. Ventura, C. Romero, A. Zafra, J. A. Delgado, C. Hervas, "JCLEC: A java framework for evolutionary computation soft computing." Soft Computing, vol. 4, no. 12, pp: 381–392, 2008.

[16]. Rong Gang, Liu Jin-feng, Gu Hai-jie, "Mining Dynamic Association Rules in Databases", Control Theory & Applications, 24(1), 2007.

[17]. Ansaf Salleb-Aouissi, Christel Vrain, Cyril Nortet "QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules", IJCAI-07

[18]. Nan Jiang and Le Gruenwald "Research Issues in Data Stream Association Rule Mining", SIGMOD Record, Vol. 35, No. 1, 2006.

[19]. S. Y. Wang, K. Tai, M. Y. Wang. "An enhanced genetic algorithm for structural topology optimization", International Journal for Numerical Methods in Engineering, 65, pp: 18-44, 2006.