A Novel Optimal Pattern Mining Algorithm using Genetic Algorithm

Vishakha Agarwal
M. Tech Scholar
Department of Computer Science & Engineering
Madhav Institute of Technology & Science, Gwalior, India

Akhilesh Tiwari
Associate Professor
Department of Computer Science & Engineering
Madhav Institute of Technology & Science, Gwalior, India

ABSTRACT

As the need for strategic information is escalating at a tremendous rate worldwide, it has become a challenging task to manage the growing data and efficiently mine useful knowledge out of this raw data. Pattern Mining is one such technique to mine useful patterns out of the pattern warehouse. Patterns are the one way of knowledge representations. Few pattern generation and mining approaches are available in literature but they are only confined to find simple patterns. Since, genetic algorithm is a search heuristic, which is used to find optimal solutions for the optimization and search problems. So in this paper, author is proposing a novel pattern mining algorithm for finding optimal patterns from pattern warehouse. The algorithm uses the features and operators of the genetic algorithm to incorporate the property of optimality among patterns.

Keywords

Data Mining, Genetic Algorithms, Pattern Mining, Pattern Warehousing.

1. INTRODUCTION

Due to recent technological advancements and affordable storage facilities, data management plays a key role in all decision making process. At every time instant, massive data is been generated in all parts of the internet. So, there has always been a need of the technology by which data generation, storage and management is performed in a well-organized manner.

1.1 Data Mining and Warehousing

Data mining plays a vital role in current growing rate of internet data. It is a process of extracting valuable information from large amount of data which is either stored in databases, data warehouses or other huge repositories. Data mining is a synonym for another quite popular term called Knowledge Discovery from Databases [1]. Some of the data mining process efficiently mine valuable patterns from the data stored in data warehouse. These patterns are the subject of interest to many business analyst in making analytical decisions for various small and big organizations.

Data warehouse [2] is one such repository which is capable of storing large amount of multidimensional data from heterogeneous sources in an integrated and persistent manner. Data warehouses acts as a source repository for performing data mining operations to generate certain analytical results i.e. extracting knowledge from the raw data. The concept of data warehouses was introduced to resolve the issue of information crises within the enterprise.

1.2 Pattern Mining and Warehousing

As, author has discussed earlier about patterns that knowledge can also be represented in the form of patterns [3] which are buried within these large repositories. Patterns can be represented as-

P1- {Milk, Bread, Butter, Eggs}

Above stated pattern P1 is an example of frequent pattern of size 4 i.e. pattern contains four items which are frequently brought together from a retail store.

Through various data mining operations, patterns can be extracted, but these patterns are volatile in nature. So a new repository is introduced called as Pattern warehouse [4] which is used to store these volatile patterns in a persistent manner and further analytical results are generated with the help of these patterns. Pattern warehouses are smaller in size than data warehouses and store more sophisticated form of information than data warehouse.

Pattern Mining is performed upon the patterns stored in pattern warehouse for generating analytical outcomes in support of business decision making process [5]. Through pattern mining business analyst has to deal with a small amount of information and for producing required reports and results.

1.3 Genetic Algorithms

Please use Genetic algorithms [6] are the search algorithms, been developed by John Holland in 1975 at university of Michigan. This heuristic mimics the mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search and optimization algorithm with some of the innovative flair of human search [7]. It uses various selection, crossover and mutation operators [8] to evolve an optimal solution which is better which is better than solutions of previous generations. The flowchart of simple genetic algorithms is shown in Figure 1.

The genetic uses three basic operators as:

Selection - This operator selects the two mating strings from old population based upon their fitness value. The higher the fitness value the more likely they are able to create new strings.

Crossover - In this operator, substrings of two parents are interchanged to create two new offsprings. The higher the crossover probability the better is the new population as it contains less strings from old population.

Mutation – This operators yields change in the new offspring with very low probability so that algorithm would not stuck and to carry out slight variation in strings. The change is observed as flipping of any bit value from 1 to 0 or vice-versa.

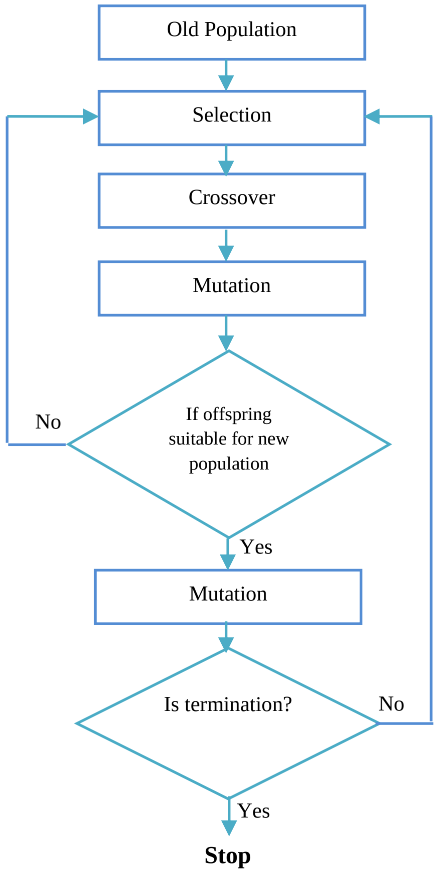

Fig 1: Workflow of Genetic Algorithm

2. PROBLEM STATEMENT

There are various data mining algorithms available in literature which generates the patterns [9] [10], but out of these patterns some of them are false frequent [11]. So, there should be some approach which removes these false frequent, and less frequent patterns and do not store them in pattern warehouse. Also, it provides some futuristic vision about frequent patterns [12].

3. PROPOSED STATEMENT

In the proposed approach, author is proposing an algorithm (OPM-GA) which works upon the optimization engine for generating optimal patterns from the pattern warehouse. The proposed algorithm uses genetic algorithm for finding optimal patterns. Figure 2 shows the scenario where the proposed algorithm is incorporated in the architecture.

Here, author presents the steps of algorithm step by step and then finally draws a flowchart and provides an example for the execution of the whole process.

Steps of algorithm (OPM-GA)

1. Select a pattern-set.
2. Enter the values of crossover probability, mutation probability and fitness value.

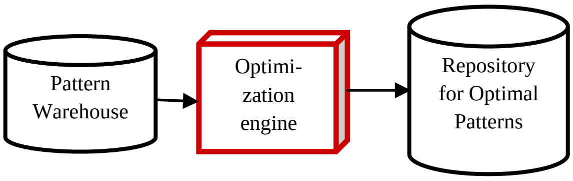

Fig 2: Workflow of Optimal Patterns

3. Apply selection operator for selection of two patterns from old population.
4. Perform crossover with given crossover probability.
5. Perform bitwise mutation.
6. Test the new offsprings (patterns) for their suitability in new population by calculating the fitness value.
7. If the pattern is suitable, place it into optimal pattern-set and further go for selection process.
8. Else, discard the pattern and go to selection process for next set of parent patterns.
9. Repeat this process until whole pattern-set is scanned.

Now, the whole process of the proposed algorithm is shown in Figure 3 in the architecture of the proposed algorithm.

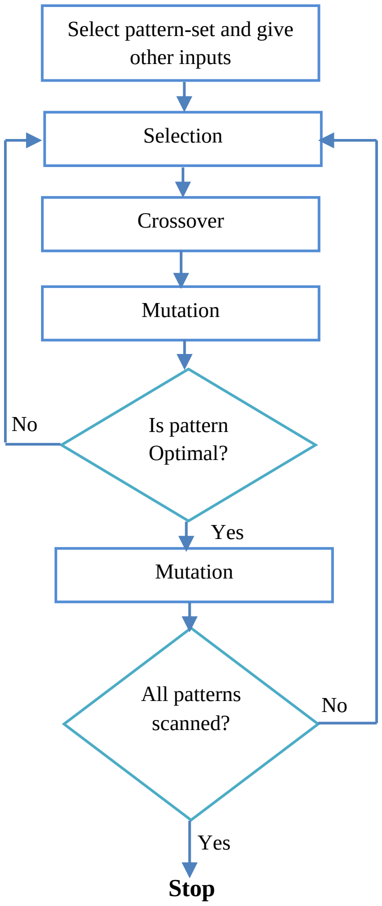

Fig 3: Flowchart of Proposed Algorithm

Demonstration of Proposed Algorithm:

Now, the author demonstrates the algorithm with the help of a sample dataset whose frequent pattern-set is considered as an input to the proposed algorithm. Table 1 describes the elements of sample dataset. In this dataset, there are six transactions of six items namely A, B, C, D, E, F.

Table 1. Sample Dataset

Transaction Id	Items
T100	A, B, C, D, E
T200	A, B, C, D, F
T300	A, B, C
T400	A, B, D
T500	A, C, D
T600	B, C, D, E

Now, extended apriori algorithm [13] [14] is applied upon this dataset with a support of 50% to extract frequent patterns which are listed in Table 2.

Table 2. Frequent Pattern-set

Pattern Id	Pattern	Support_value
P1	A	5
P2	B	5
P3	C	5
P4	D	5
P5	A, B	4
P6	A, C	4
P7	A, D	4
P8	B, D	4
P9	B, C	4
P10	C, D	4
P11	A, B, C	3
P12	A, B, D	3
P13	A, C, D	3
P14	B, C, D	3

Now, the pattern-set shown in Table 2 is given as an input to the algorithm OPM-GA along with few other inputs as crossover probability is taken as 95% and mutation probability is taken as 1%. The algorithm works as follows-

Selection- In the selection process, the two patterns are randomly selected. For eg.

P1- A (Parent 1)

P14- B, C, D (Parent 2)

Crossover- Now, crossover operator is applied over these two parent patterns and two new offsprings are generated.

NP1- A, D (Offspring 1)

NP2- B, C (Offspring 2)

Mutation- After crossover, these two new offsprings are mutated by a single bit and further tests for their fitness in the new population. If the patterns pass their fitness value these are placed in the new population of optimal patterns. Moreover, same process is repeated for other set of parent patterns. Table 3 shows the optimal patterns extracted out of the old population of pattern-set.

Table 3. Optimal Pattern-set

Pattern Id	Pattern
P1	A
P2	B
P5	A, B
P10	C, D
P11	A, B, C
P12	A, B, D
P13	A, C, D
P14	B, C, D

4. RESULT ANALYSIS

After implementing the whole algorithm over the sample data-set, author discovered that initially, the number of frequent patterns that were stored in pattern warehouse was 14 but when genetic based proposed algorithm is applied i.e. OPM-GA, and whole pattern-set is passed through the optimization engine this number has narrowed down to 8 patterns which are found to be the optimal ones i.e. the patterns that do not fall under the category of false frequent patterns or less frequent ones when database gets updated. This analysis is shown below with the help of a graph in Figure 4.

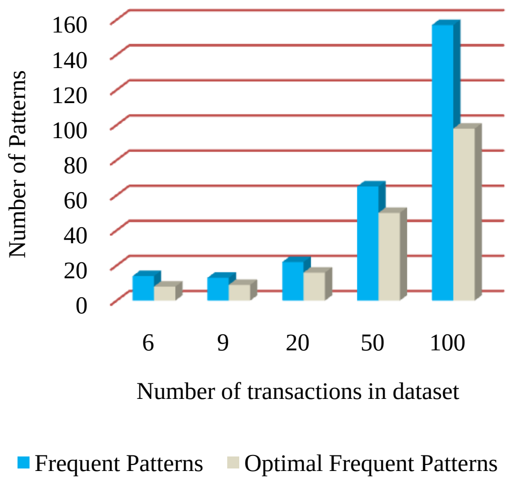

Fig 4: Graph showing reduction in no. of patterns after applying OPM-GA on varying sized datasets

Here, Figure 4 shows a comparative analysis which describes the reduction in number of patterns that are extracted on applying the proposed algorithm on different datasets having varying number of transactions.

5. CONCLUSION

A great deal of work has been done over genetic based frequent pattern mining and genetic based association rule mining. But, since pattern warehousing is a new concept, very few researchers has looked into this concept of making the patterns non-volatile and store them in persistent manner. The approach proposed by the author is one step ahead of just finding the patterns, efficiently store those patterns and manage the pattern warehouse. The algorithm generates and stores optimal patterns which helps in the further reduction of the size of the repository. The inclusion of genetic algorithm provides the strength to the proposed approach. The method described here is very simple and efficient one. In future, work can be extended by incorporating various other heuristic approaches and defining other new measures in order to find more interesting patterns that are applicable to applications related to different domains.

6. REFERENCES

[1] Dunham, M. H. 2006 Data Mining: Introductory and Advanced Topics. Pearson Education.

[2] Ponniah, P. 2001 Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. Wiley India.

[3] Bartolini, I., Bertino, E., Catania, B., Ciaccia, P., Golfarelli, M., Patella, M. and Rizzi, S. 2003 Patterns for Next-generation Database Systems: Preliminary Results of the PANDA Project. In Proceedings of the Eleventh Italian Symposium on Advanced Database Systems, SEBD, Cetraro (CS), Italy.

[4] Agarwal, V. and Tiwari, A., “From Data Warehouse to Pattern Warehouse: A Progressive Step”, International Journal of Engineering Research”, 2016, Vol. 5, No.4, pp: 249-252.

[5] Tiwari, V. and Thakur, R. S. “P2ms: A Phase-wise Pattern Management System for Pattern Warehouse”, International Journal of Data Mining, Modeling and Management, Inderscience, 2014, Vol. 5, No. 3, pp: 1-10.

[6] Kanani, D. S. and Mishra, S. K. “An Optimized Association Rule Mining using Genetic Algorithm”, International Journal of Computer Applications, 2015, Vol. 119 No. 14, pp:11-15.

[7] Goldberg, D. E. 2006 Genetic Algorithms in Search, Optimization and Machine Learning. Pearson Education.

[8] Verma, G. and Verma, V. “Role and Applications of Genetic Algorithm in Data Mining”, International Journal of Computer Applications, 2012, Vol. 48 No.17, pp: 5-8.

[9] Flockhart, I. W. and Radcliffe, N. J. 1996 A Genetic Algorithm-Based Approach to Data Mining, AAAI: Knowledge Discovery and Data Mining, Portland, Oregon.

[10] Marmelstein, R. E. 1997 Application of Genetic Algorithms to Data Mining. In Proceedings of Modern Artificial Intelligence and Cognitive Science, Dayton, Ohio.

[11] Tiwari, A., Gupta, R. K. and Agrawal, D. P., “A Survey on Frequent Pattern Mining: Current Status and Challenging Issues”, Information Technology Journal, 2010, Vol. 9, No.7, pp: 1278-1293.

[12] Agrawal, R., Imielinski, T. and Swami, A. 1993 Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of data, New York, USA. pp: 207-216.

[13] Pujari, A. K. 2007 Data Mining Techniques. Universities Press.

[14] Han, J. and Kamber, M. 2006 Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.