A Review on Mining High Utility Itemsets

Sruthy John[1], Dr. Anithakumari[2]

[1]*M.Tech Student, Department of Computer Science and Engineering, LBS Institute of Technology for Women Thiruvananthapuram, Kerala, India*
[2]*Professor, Department of Computer Science and Engineering, LBS Institute of Technology for Women Thiruvananthapuram, Kerala, India*

Abstract - *Data mining is the process of mining useful information from a database. ARM finds items that are brought together. FIM find out items that are frequent in a database. These two mining techniques do not consider the quantity and profit of the items purchased. It is addressed in High Utility Itemset Mining(HUIM) . Thus HUIM came into existence, that mines the items from the transaction database which generates high profit. This paper targets on reviewing existing state of art algorithms and to provide a path for future research in the field of high utility itemset mining.*

Key Words: Association Rule Mining, Frequent Itemset mining, High Utility, Transaction Database.

1. INTRODUCTION

The goal of frequent itemset mining is to find frequent itemsets. Many popular algorithms have been proposed for this problem such as Apriori, FPGrowth, LCM, Eclat, etc. These algorithms takes as input a transaction database and a parameter “minsup” called the minimum support threshold. These algorithms then return all set of items (itemsets) that appears in at least minsup transactions. An important limitation of frequent itemset mining is that purchase quantities are not taken into account. Thus, an item may only appear once or zero time in a transaction. Thus, if a customer has bought five breads, ten breads or twenty breads, it is viewed as the same. All items are viewed as having the same importance, utility of weight. Thus, frequent pattern mining may find many frequent patterns that are not interesting. In utility mining, each item has its own profit and can occur multiple times in one transaction. The utility of an itemset is calculated by summing the product of the item’s profit and its occurrence quantity in each relevant transaction. High utility itemsets (HUIs) are those whose utility is no lower than a user-specified threshold. The problem of high utility itemset mining (HUIM) is to discover all HUIs within a transaction database. Various mining algorithms have been proposed for the discovery of HUIs. The existing exact approaches for HUIM tend to degrade as the size of the database and the number of distinct items increase, and the performance may become unacceptable, similar to the problem of FIM applied to social networks or large bioinformatics datasets. To deal with the performance bottleneck of exact approaches, bio-inspired algorithms have been applied for HUIM. HUIM is different from problems in which there are relatively few best values—all itemsets with utilities no lower than the minimum threshold must be discovered. Because the distribution of HUIs is not even, searching with the best values from the previous population as targets may mean that some results are missed within a certain number of iterations. To solve this problem, a novel bio-inspired-algorithm-based HUIM framework(Bio-HUIF) based on bat algorithm has been developed to discover HUIs. Bat inspired algorithm is a meta heuristic optimization algorithm developed by Xin-She Yang in 2010.This algorithm is based on the echolocation behavior of micro bats with varying pulse rates of emission and loudness.

1.1 Frequent Itemset Mining

The real motivation of finding frequent sets is to analyze supermarket transaction database[9], so as to examine the customer behavior in terms of the purchased data. It aims at finding regularities in the shopping behavior of customers of supermarkets, mail-order companies, on-line shops etc. The process of discovering all frequent itemsets from transactional database is quite difficult. As the search space is exponential in the number of items occurring in the database and the targeted database tend to be massive, which contains millions of transactions. Therefore most efficient techniques are needed to solve this task. The possible applications of FIM are:

- Helps in arrangement of products in shelves on a catalog’s pages etc.
- Suggesting other products, product bundling.
- Fault localization, fraud detection, technical dependence analysis etc.

1.2 Association Rule Mining

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis. The researches on frequent itemset mining identified that infrequent itemsets provides useful insight into the dataset. Therefore a knowledge discovery problem called indirect association has been proposed[10]. An indirect weighted association rule mining model is an extension of indirect association rule mining, which states the problem that downwards closure property is invalid in the weighted association rule mining. model. A factor is set which relates minimum support with weighted minimum support so as to maintain a property that if an itemset is a weighted frequent itemset under the weighted minimum support then the itemset must be a frequent itemset under the weighted minimum support.

Therefore, we can firstly find frequent K-itemset L_K, then generate weighted K-frequent itemset WL_K from L_K.

2. Related Works

GA is an evolutionary based stochastic optimization algorithm with a global search potential proposed by Holland in 1975[11]. GA adapts the principles of Charles Darwin Theory of survival of the fittest. Genetic algorithm begins by initializing a population of solution (chromosome). It comprises representation of the problem usually in the form of a bit vector. Then for each chromosome evaluate the fitness using an appropriate fitness function suitable for the problem. Based on this ,the best chromosomes are selected into the mating pool, where they undergo cross over and mutation thus giving new set of solutions(offspring)

GA is useful and efficient when:

- The search space is large
- No mathematical analysis is available.
- Domain knowledge is less to encode to narrow the search space
- For complex or loosely defined problems since it works by its own internal rules.

GA has some disadvantages:

- If the fitness function is not defined properly it mayhave a tendency to converge towards local optima rather than the global optimum of the problem.
- Operating on dynamic data sets is difficult.
- For specific optimization problems, and given the same amount of computation time, simpler optimization algorithms may find better solutions than GAs.
- GAs are not directly suitable for solving constraint optimization problems.

Evolutionary programming[12] is a global optimization algorithm inspired by the theory of adaptation and evolution by means of natural selection. Unlike GA which deals with micro or genomic level, this technique deals with macro-level or species level process of evolution. ES utilizes self-adaptive mechanisms for controlling the application of mutation.

Paddy field algorithm[13] operates on a reproductive principle. Unlike evolutionary algorithms ,it does not involve combined behavior nor crossover between individuals instead it uses pollination and dispersal.

The basic five steps in PFA are:

1. Sowing: The algorithm operates by initially scattering seeds (initial population p0) at random in an uneven field.
2. Selection: Here the best plants are selected based on a threshold method so as to selectively weed out unfavorable solutions and also controls the population.
3. Seeding: In this stage each plant develops a number of seeds proportional to its health. The seeds that drop into the most favorable places (most fertile soil, best drainage, soil moisture etc.) tend to grow to be the best plants (taller) and produce more number of seeds. The highest plant of the population would correspond to the location of the optimum conditions and the plant‘s fitness is determined by a fitness function.
4. Pollination: For seed propagation pollination is a major factor either via animals or through wind. High population density would increase the chance of pollination for pollen carried by the wind.
5. Dispersion: In order to prevent getting stuck in local minima, the seeds of each plant are dispersed. Depending on the status of the land it will grow into new plants and continue the cycle.

Ant colony optimization is a successful swarm based algorithm[14]. It uses the foraging behaviour of ants known as stigmergy. The most interesting aspect is the ability of the ants to find the shortest path between the ants nest and the food source by tracing the pheromone trails. Ants deposit pheromones on the paths they are following this phenomenon results in self-reinforcing process which creates a path that is composed of high pheromone concentration. ACO is used for solving complex optimization problems.

There are mainly three functions structured in ACO:

1. AntSolutions Construct: It performs the solution construction process where the artificial ants move through adjacent states of a problem according to a transition rule, iteratively building solutions.
2. Pheromone Update: performs pheromone trail updates. This may involve updating the pheromone trails once complete solutions have been built, or updating after each iteration. In addition to pheromone trail reinforcement, ACO also includes pheromone trail evaporation. Evaporation of the pheromone trials helps ants to forget bad solutions that were learned early in the algorithm run.
3. Deamon Actions: is an optional step in the algorithm which involves applying additional updates from a global perspective (for this no natural counterpart exists). This may include applying additional pheromone reinforcement to the best solution generated.

Yadav et al.[3] proposed a data structure for finding maximum frequent item set in online data mining. The data structure consists of a tree, which is known as Ordered Tree. It is a one pass algorithm. The Ordered Tree for locating maximum frequent Itemset consist of 26 path for every alphabetical letter, thus it also known as multipath tree. Structure of a Ordered Tree such that label of root node is null and the child node start sequentially and end with letter

Z. Node of every path decreases one by one with the increasing path of a Ordered Tree. Finding maximum frequent item set in online data mining the ordered tree works as follows. Apply the sorting in each transaction. After that, sort element insert into Ordered Tree according to their alphabetical order if encoded in alphabets. In a Ordered Tree same element of a different path connected to each other, so finding frequent itemsets will be an easier process. Applying sorting to every transaction. Insert transaction, one by one into Ordered Tree depend upon its prefix structure. When same value will be repeated, increment the value by one in Ordered Tree. This is an approximation-based approach. The data structure used in this method overcome the problem of unnecessary scanning process.

The challenges faced by data analyst are search space for mining HUIs is increasing, the size of the search space is determined by the size of the transaction and the number of distinct items in a transaction database[1]. Data analyst need to specify the minimum threshold to mine HUI. Genetic algorithm based techniques are designed to mine HUI from the transaction database. This approach can handle large number of distinct items and transaction. It works well on itemsets containing negative item values. But the number of HUIs discovered is less when the transaction database is huge.

Tsang et al.[2] proposed that the situation may become worse when the database contains lots of long transactions or long high utility itemsets. Two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility itemsets with a set of effective strategies for pruning candidate itemsets. The information of high utility itemsets is maintained in a tree-based data structure named utility pattern tree (UP-Tree) such that candidate itemsets can be generated efficiently with only two scans of database. The performance of UP-Growth and UP-Growth+ is compared with the state-of-the-art algorithms on many types of both real and synthetic data sets. A discrete PSO-based algorithm, namely HUIM-BPSO, is designed to find the HUIs by integrating the sigmoid updating strategy[4] and TWU model. An OR/NOR-tree structure [5] is developed to reduce the multiple database scans by early pruning the invalid combinations of the particles. This process can greatly reduce the computations of the invalid particles. But it cannot handle the issues of HUIM.

In addition to genetic algorithm, there are many other bio-inspired approaches, which are designed to extract association rules. In [6] G3PARM algorithm is developed, it is based on genetic programming. The authors used grammar guided genetic programming (G3P) to avoid invalid individuals found by Genetic Programming(GP) process. Also, G3PARM permits multiple variants of data by using a context free grammar

3. CONCLUSION

From the survey done it came to know that bio-inspired approaches outperforms well in terms of convergence speed and optimization than the state of art algorithms used for data mining. Most of the state of art algorithms used for data mining hat to maintain some data structures for storing and processing the data, while bio inspired algorithms are not maintaining any data structure as they are mimicking the biological behavior of organisms and sub-organisms for solving the optimization problem. In future bio-inspired algorithms could be hybridized with each other methods and approaches so as to enhance the performance of bio-inspired optimization algorithms.

REFERENCES

[1] S. Kannimuthu & K. Premalatha (2014) “Discovery of High Utility Itemsets Using Genetic Algorithm with Ranked Mutation”, Applied Artificial Intelligence, 28:4, 337-359.

[2] V. S. Tseng, B. Shie, C. Wu and P. S. Yu, "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases," in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 8, pp. 1772-1786, Aug. 2013.

[3] L. Yadav and P. S. Nair, "A new data structure for finding maximum frequent itemset in online data mining," 2015 International Conference on Computer, Communication and Control (IC4), Indore, 2015, pp. 1-5.

[4] W. Liu, Z. Wang, Y. Yuan, N. Zeng, K. Hone and X. Liu, "A Novel Sigmoid-Function-Based Adaptive Weighted Particle Swarm Optimizer," in IEEE Transactions on Cybernetics.

[5] Lin, J.C., Yang, L., Fournier-Viger, P. et al. A binary PSO approach to mine high-utility itemsets. Soft Comput 21, 5103–5121,2017.

[6] Olmo Ortiz, Juan Luis & Luna, José María & Romero, José Raúl & Ventura, Sebastian. (2011). Association Rule Mining using a Multi-Objective Grammar-Based Ant Programming Algorithm. International Conference on Intelligent Systems Design and Applications.

[7] Jin Gou, Fei Wang & Wei Luo (2015) Mining Fuzzy Association Rules Based on Parallel Particle Swarm Optimization Algorithm, Intelligent Automation & Soft Computing, 21:2, 147-162.

[8] Wang, Y.; Wang, P.; Zhang, J.; Cui, Z.; Cai, X.; Zhang, W.; Chen, J. A Novel Bat Algorithm with Multiple Strategies Coupling for Numerical Optimization. Mathematics 2019, 7, 135.

[9] L. Yadav and P. S. Nair, "A new data structure for finding maximum frequent itemset in online data mining," 2015 International Conference on Computer, Communication and Control (IC4), Indore, 2015, pp. 1-5.

[10] W. Ouyang and Q. Huang, "Discovery Algorithm for Mining both Direct and Indirect Weighted Association Rules," 2009 International Conference on Artificial Intelligence and Computational Intelligence, Shanghai, 2009, pp. 322-326.

[11] J.H. Holland, Genetic algorithms and the optimal allocation of trials, SIAM J. Comput. 2 (2) (1973) 88–105.

[12] Beyer, H.G. and Schwefel, H.P. 2002: Evolution strategies. Natural Computing 1,3–52.

[13] Upeka Premaratne , Jagath Samarabandu, and Tarlochan Sidhu, ―A New Biologically Inspired Optimization Algorithm‖,Fourth International Conference on Industrial and Information Systems, ICIIS 2009,28-31 December 2009, Sri Lanka.

[14] Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant System: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics – Part B, 26, 29–41.