International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 13855-13862

http://www.ripublication.com

Optimizing The Achieved Frequent Item Sets Using Genetic Algorithm

Paul P Mathai
Research Scholar at Noorul Islam University, Tamil Nadu cum Asst. Professor at Federal Institute of Science and Technology, Kerala
R.V. Siva Balan
Associate Professor, Noorul Islam University, Tamil Nadu

Abstract

Frequent pattern mining casts a vital part in many significant data mining task such as associations, sequential patterns, partial periodicity, to name a few. Nevertheless, it is common knowledge that frequent pattern mining habitually produces an awfully mammoth number of frequent item sets and rules, paving the way for diminution in competence as well as efficacy of extraction in view of the fact that clients are best with the task of sieving through a huge number of extracted rules to locate the fruitful ones. Therefore, without resorting to the extraction of the frequent itemsets, mining only closed frequent itemsets goes a long way in incredibly enhancing the excellence along with the cutback in the calculation period. In the innovative closed frequent item set mining, the data are gathered from the database and the support of each and every itemset is calculated by moving the sliding window. Subsequently, the frequent itemsets are extracted by demarcating a pre-fixed threshold. The achieved frequent itemsets are thereafter furnished to the Genetic Algorithm (GA) for the purpose of optimization. The method is executed in the JAVA platform and the excellence of the well-conceived approach is analyzed and contrasted with the modern methods.

Keywords: Data mining, Knowledge Discovery Database (KDD), Itemsets, Frequency

Introduction

The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer based methodology, including new techniques, for discovering knowledge from data is called data mining [1]. The objective of data mining is to identify valid novel, potentially useful, and understandable correlations

and patterns in existing data. Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing). The term "data mining" is primarily used by statisticians, database researchers, and the MIS and business communities. The term Knowledge Discovery in Databases (KDD) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process [2] [3]. Data mining is a highly inter disciplinary area spanning a range of disciplines; statistics, machine learning, databases, pattern recognition and other areas [4].

In general, data mining methods can be classified into two categories; predictive and descriptive. Predictive data mining methods predicts the values of data, using some already known results that have been found using a different set of data. Predictive data mining tasks include: classification, prediction. Descriptive mining tasks characterize the general properties of the data in database. This is done by identifying the patterns and relationships in the data [5]. In data mining, items are mined from the database based on two constraints: items frequency and utility.

Calculating itemset support (or frequency counting) is a fundamental operation that directly impacts space and time requirements of many widely used data mining algorithms. Some data mining algorithms (i.e., frequent itemset mining) are only concerned with identifying the support of a given query itemset, while others (i.e., pattern-based clustering algorithms) must in addition identify the transactions that contain the query itemset [6]. The goal of frequent itemset mining is to find items that co-occur in a transaction database above a user given frequency threshold, without considering the quantity or weight such as profit of the items. However, quantity and weight are significant for addressing real world decision problems that require maximizing the utility in an organization. The high utility itemset mining problem is to find all itemsets that have utility larger than a user specified value of minimum utility [7] [8].

Utility-based data mining is a broad topic that covers all aspects of economic utility in data mining. It encompasses predictive and descriptive methods for data mining, among the later especially detection of rare events of high utility (e.g. high utility patterns) [9]. Utility based data mining refers to allowing a user to conveniently express his or her perspectives concerning the usefulness of patterns as utility values and then finding patterns with utility values higher than a threshold. A pattern is of utility to a person if its use by that person contributes to reaching a goal [10].

Related Work

Identifying the association rules in large databases play a key role in data mining. Kalli Srinivasa Nageswara *et al.* [11] have considered the prior researches and present working status in order to restore the gaps between them with present known information. There were two problems regarding this context: identifying all frequent item sets and to generate constraints from them. Here, first problem, as it takes more processing time, was computationally costly. Consequently, many algorithms were

proposed to solve this problem. Their current study considers such algorithms and the related issues.

Saravanabhavan *et al.* [12] have presented an efficient tree structure for mining high utility itemsets. At first, they have developed a utility frequent-pattern tree structure, an extended tree structure for storing crucial information about utility itemsets. Then, the pattern growth methodology was utilized for mining the complete set of utility patterns. Improved high utility itemsets mining efficiency was achieved using two major concepts: 1) Compressing a large database into a smaller data structure as well as the utility FP-tree avoids repeated database scans, 2) The pattern growth method utilized in the proposed FP-tree-based utility mining avoids the costly generation of a large number of candidate sets and thereby reduces the search space dramatically. Experimental analysis was carried out on tree structure mining concept using different real life datasets. The performance evaluation results have demonstrated the efficiency of the proposed approach in mining high utility itemsets.

Association Rules are the most important tool to discover the relationships among the attributes in a database. Vijaya Prakash *et al.* [13] have discussed that the existing Association Rule mining algorithms were applied on binary attributes or discrete attributes, in case of discrete attributes there was a loss of information and these algorithms take too much computer time to compute all the frequent itemsets. By using Genetic Algorithm (GA), it is possible to improve the generation of Frequent Itemset for numeric attributes. The major advantage of using GA in the discovery of frequent itemsets is that they perform global search and its time complexity was less compared to other algorithms as the genetic algorithm was based on the greedy approach. The main aim of their paper is to find all the frequent itemsets from given data sets using genetic algorithm.

The main goals of Association Rule Mining (ARM) are to find all frequent itemsets and to build rules based of frequent itemsets. But a frequent itemset only reproduces the statistical correlation between items, and it does not reflect the semantic importance of the items. To overcome this limitation, Kannimuthu *et al.* [14] have utilized a utility based itemset mining approach. Utility-based data mining is a broad topic that covers all aspects of economic utility in data mining. It takes in predictive and descriptive methods for data mining. High utility itemset mining is a research area of utility based descriptive data mining, aimed at finding itemsets that contribute most to the total utility. The well known faster and simpler algorithm for mining high utility itemsets from large transaction databases is Fast Utility Mining (FUM). In this proposed system, they made a significant improvement in FUM algorithm to make the system faster than FUM. The algorithm was evaluated by applying it to IBM synthetic database. Experimental results have shown that the proposed algorithm was effective on the databases tested.

Parvinder S. Sandhu *et al.* [15] have proposed an efficient approach based on weight factor and utility for effectual mining of significant association rules. Initially, the proposed approach has utilized traditional Apriori algorithm to generate a set of association rules from a database. The proposed approach exploits the anti-monotone property of the Apriori algorithm, which states that for a k-itemset to be frequent all (k-1) subsets of this itemset also have to be frequent. Subsequently, the set of

association rules mined were subjected to weightage (W-gain) and utility (U-gain) constraints, and for every association rule mined, a combined utility weighted score (UW-Score) was computed. Ultimately, they have determined a subset of valuable association rules based on the UW-Score computed. The experimental results have demonstrated the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development.

Venu Madhav Kuthadi [16] has proposed an enhanced association rule mining algorithm to mine the frequent patterns. The algorithm utilized weightage validation in the conventional association rule mining algorithms to validate the utility and its consistency in the mined association rules. The utility is validated by the integrated calculation of the cost/price efficiency of the itemsets and its frequency. The consistency validation is performed at every defined number of windows using the probability distribution function, assuming that the weights are normally distributed. Hence, validated and the obtained rules are frequent and utility efficient and their interestingness are distributed throughout the entire time period. The algorithm was implemented and the resultant rules were compared against the rules that can be obtained from conventional mining algorithms

Proposed Methodology

In our ambitious frequent item set optimization technique, the data are gathered from the database. Subsequently, support of each and every itemset is calculated by moving the sliding window. Thereafter, the frequent itemsets are extracted by demarcating a pre-fixed threshold [17]. The achieved frequent itemsets are afterward furnished to the Genetic Algorithm (GA) for the sake of optimization.

Frequent Itemset Mining

Let D be the data stream comprising a sequence of transactions. Each and every transaction is linked with an identifier, named TID and therefore the transactions are characterized as $TID_1, TID_2, \ldots TID_n$, where, n goes on fluctuating for an indefinite period as the data stream is indefinite. Each and every transaction is a compendium of items, labeled as itemsets signified as $TID_1 = I_{m1}, I_{m2}, \ldots I_{mn}$, where, I_m stands for the item i.e. $I_m \subseteq \{I\}$ and $|I_{m1}, I_{m2}, \ldots I_{mn}| \leq |I|$.

At the outset, a window with specific dimension of transactions is mined from the gathered data stream. The sliding window task is carried out by supplementing one transaction and doing away with the time-barred one. With the deft deployment of traditional a-priori algorithm, the association rules are mined in accordance with the specified minimum support S_T (the rules are mined according to the user-desired; minimum support value) and confidence value. Thereafter, the subsequent window of transactions is mined followed by the extraction of the rules. The support for the rules are estimated by means of Equation 1 furnished below:

$$S = ff_I / W_s \tag{1}$$

Where, W_s and ff_I represent the window dimension and frequency of the itemset window correspondingly. The window dimension is selected according to the data stream dimension. From each and every window closed frequent item sets are extracted with the assistance of the support and hash table data. In addition to the support, the closed itemsets are chosen taking into consideration its utility value along with the stability of the utility for a fixed number of windows of transactions.

Optimization of Frequent Itemset Using Genetic Algorithm

Prior to the extraction of the closed frequent itemsets, the frequent itemsets gathered from each and every window are optimized by means of the Genetic Algorithm with the intention of keeping at bay the erroneously extracted frequent itemsets. In this regard, Genetic algorithm is an outstanding meta-heuristic algorithm which is equipped with the prowess of diminishing the innate evolution procedure. Habitually, it is employed to generate profitable solutions to optimization and search issues. It creates solutions to optimization dilemmas by means of employing methods triggered by ordinary evolutions like inheritance, mutation, selection, and crossover.

Initial Phase- At the outset, the populations of the chromosomes x_i,($i=1,2,..N,$) are created discretely. N signifies the dimension of the population. The chromosomes are home to the frequent itemsets gathered from the sliding window function.

Fitness function- Fitness value of each and every constraint is ascertained and the chromosome possessing the maximum fitness value is shortlisted as the finest chromosome. At this juncture, the fitness is estimated in accordance with Equation 2 shown below:

$$F_i=\max(f_i) \tag{2}$$

$$f_i=\sum_{i=1}^{|I_i|} freq_i \tag{3}$$

Equation. (3) Indicates the aggregate of the frequency of an itemset.

Mutation- In the course of the mutation function, chromosome values are altered in accordance with the probability. The probability to work out the mutation task has to be fixed at a lower level.

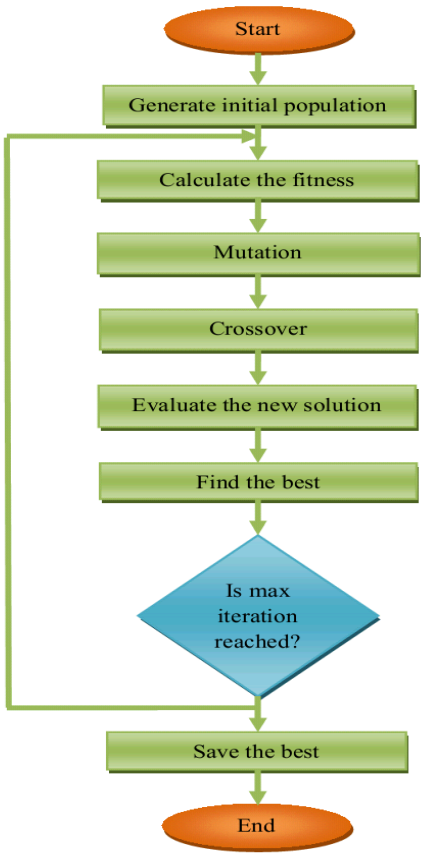

Figure 1: Flowchart for GA

Crossover- When mutation is over, one or more parent chromosomes are chosen by means of roulette wheel and fresh solution is generated.

Evaluation- Subsequently, the fitness of the fresh solution is ascertained with the aid of Equation (2). If the fresh fitness value is superior to the existing fitness value, it is substituted in place of the existing value. The procedure gets replicated till attainment of the stoppage benchmark.

Results and Discussions

The performance of our proposed frequent item set mining technique is analyzed by using the TCET database, where our proposed frequent itemset mining technique's performance is evaluated and compared with the other frequent itemset mining techniques.

The first main objective of the proposed technique is reducing the computation time while extracting FI's. The proposed technique has taken low computation time than the other techniques. While increasing the support value, the computation time is decreasing. By achieving the low computation time our proposed technique has met its first objective.

When comparing the computation time of all the techniques, our proposed FI technique has taken low computation time even though the computation time increases as the size of the window increases. One of the main objectives of the proposed technique is obtaining low memory usage or reducing the memory usage. When compared to memory usage of the other techniques, the proposed technique consumes very low memory space.

Conclusions

In this paper, we have proposed a Frequent Itemset mining algorithm for mining frequent item sets from the data streams. The proposed technique is implemented in the JAVA platform. The computation of frequent item sets from the data stream minimizes the memory usage and processing time. Thus our proposed technique performance is analyzed by using the database and compared with the FI and FI without GA techniques. The obtained results of our proposed and existing mining techniques performance in terms of run time and memory usage obtained have shown that our proposed Closed Frequent Itemset mining technique has consumed only a smaller amount of run time and a higher number of frequent itemset than the other techniques.

References

[1]. Neda Khalilzadeh and Parham Jafari Moghadam Fard, "Application of Data Mining in Marketing and Managing Customer Relationship", In Proceedings of the Marketing Management Conference, pp. 1-13

[2]. Joyce Jackson, "Data Mining: A Conceptual Overview", Communications of the Association for Information Systems, Vol. 8, pp. 267-296, 2002

[3]. Usama Fayyad, Gregory Piatetsky - Shapiro and Padhraic Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, No. 11, 1996

[4]. J Deogun, V Raghavan, A Sarkar, H Sever, "data mining: Research Trends, challenges, and applications", In Roughs Sets and Data Mining Analysis of Imprecise Data, pp. 9-45, 1997

[5]. Slavco Velickov and Dimitri Solomatine, "Predictive Data Mining: Practical Examples", In Proceedings of 2nd workshop on Artificial Intelligence in Civil Engineering, Cottbus, Germany, pp. 1-16, 2000

[6]. Hassan H. Malik and John R. Kender, "Optimizing Frequency Queries for Data Mining Applications", In Proceedings of the Seventh IEEE International Conference on Data Mining, pp. 595-600, 2007

[7]. Alva Erwin, Raj P. Gopalan and Achuthan, "Efficient Mining of High Utility Itemsets from Large Datasets", In Proceedings of PAKDD, Osaka, Japan, 2008

[8]. Alva Erwin, Raj P. Gopalan and Achuthan, "A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets", In Proceedings of the Confertence on AIDM, Gold Coast, Australia, Vol. 84, pp. 3-11, 2007

[9]. Vid Podpecan, Nada Lavrac and Igor Kononenko, "A Fast Algorithm for Mining Utility-Frequent Itemsets", In Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases, 2007

[10]. Hong Yao, Howard J. Hamilton and Liqiang Geng, "A Unified Framework for Utility Based Measures for Mining Itemsets", In Proceedings of the second Workshop on Utility-Based Data Mining (UBDM), pp. 28-37, 2006

[11]. Kalli Srinivasa Nageswara Prasad and Ramakrishna, "Frequent Pattern Mining and Current State of the Art", International Journal of Computer Applications, Vol. 26, No. 7, pp. 33-39, 2011

[12]. Saravanabhavan and Parvathi, "Utility FP-Tree: An Efficient Approach to Mine Weighted Utility Itemsets", European Journal of Scientific Research, Vol. 50 No. 4, pp. 466-480, 2011

[13]. Vijaya Prakash, Govardhan and Sarma, "Mining Frequent Itemsets from Large Data Sets using Genetic Algorithms", IJCA-Artificial Intelligence Techniques - Novel Approaches & Practical Applications, No. 4, Vol. 7, pp. 38-43, 2011

[14]. Kannimuthu, Premalatha and Shankar, "iFUM - Improved Fast Utility Mining", International Journal of Computer Applications, Volume 27– No.11, pp. 32-36, 2011

[15]. Parvinder S. Sandhu, Dalvinder S. Dhaliwal and Panda, "Mining utility-oriented association rules: An efficient approach based on profit and quantity", International Journal of the Physical Sciences Vol. 6, No. 2, pp. 301-307, 2011

[16]. Venu Madhav Kuthadi, "A New Data Stream Mining Algorithm for Interestingness-Rich Association Rules", Journal of Computer Information Systems, pp. 14-27, 2013

[17]. Paul P Mathai, R. V. Siva Balan, “An Extensive Review of Significant Researches in Data Mining”, Research Journal of Applied Sciences, Engineering and Technology Vol. 7 No. 22, pp. 4779-4794, 2014