International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Review Article

Optimization Techniques in Association Rule Mining: A Review

Parmjeet Kaur[Á*], Usvir Kaur[Á] and Dheerendra Singh[Ḃ]

[Á]Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India
[B]Shaheed Udham Singh College of Engineering and Technology, Tangori, Punjab, India

Accepted 06 May 2014, Available online 01 June2014, **Vol.4, No.3 (June 2014)**

Abstract

Current World Wide Web is featured by abundance of data. Association Rule mining algorithms dealt with data in competent manners and with practical time. Association Rule mining focuses on extracting frequent pattern association or casual structure from the given transactional database. Association Rule mining is most used to find the frequent itemsets from large database. This paper elaborates upon the use of association rule mining in extracting patterns that occurs frequently within a dataset and showcases some of the optimization techniques used for association rule mining.

Keywords: *Association Rule Mining, Bees Swarm Optimization, Tabu Search, Ant Colony Optimization.*

Introduction

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as clustering, classification, association rule mining and sequence detection (J. Vaidya *et al*, 2002). Association Rule Mining is one of the most important and well studied techniques of data mining tasks (Han, J *et al*, 2006). Mining association rules from a large database, has been an important task in the area of data mining to discover hidden, interesting associations that occur between various data items. Nowadays ARM is broadly used in many different areas such as telecommunication networks, market and risk management, inventory control mobile mining, graph mining, educational mining, etc. The pattern and rule discovered are based on the majority of commonly repeated items in dataset. Consider a transaction database, where each transaction is a set of items, and an association rule reveals the relationship between items (Kanimozhi *et al*, 2009, C. Romero et.al).
For example consider a supermarket with the large collection of items. Management decision of supermarket includes items for sale, designing of coupons, maximum profit with respect to merchandised. Previous transactions used for improvements of decision. Bar codes helps to store the items purchased on per-transaction basis called as *basket* (R. Agarwal *et al*, 1993).

Formal statement of association rule mining:

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items. Let T be a set of transactions (the database), where each transaction t (a data case) is a set of items such that $t \subseteq I$. An association rule is an Implication of the form, $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$. The rule $X \rightarrow Y$ holds in the transaction set T with confidence c if c% of transactions in T that support X also support Y. The rule has support s in T if s% of the transactions in T contains $X \cup Y$ (Liu, B *et al*, 1999). The problem of mining association rules is to generate all association rules is to generate all association rules from the given database D that have support and confidence greater than the user-specified minimum support called as *minsup* and minimum confidence called as *minconf* respectively. Our decision is neutral with respect to the representation of D. D could be data file, relational table or the result of relation expressions(R. Agarwal *et al*, 1994).

Literature Review

(Djenouri *et al*, 2013) proposed a new hybrid algorithm called (HBSO-TS) for association rule mining based on two metaheuristis algorithms Bees Swarm Optimization and Tabu Search. BSO will browse the search space in such away to cover most of its regions and the local exploration of each bee is computed by tabu search. Observation that the developed approach yields useful association rules in a reasonable time when comparing it with previous works but also found difficulties to maintain the empirical parameters settings. Computation on General Processor Unit (GPU) is planned for gain more in time.

(Djenouri *et al*, 2012) presents a new algorithm based on Genetic metaheuristic and Bees Swarm Optimization. This approach is capable with handling of large amount of data as some of existing algorithms are not capable to do so in reasonable time. Results show that concerning the fitness criterion, BSO-ARM achieved slightly better than all the genetic approaches. On the other hand, BSO-ARM is more time consuming. The two important operations

*Corresponding author: **Parmjeet Kaur**

(determination search area and neighbourhood search) provided by BSO, permit to improve the solution quality but it requires a considerable computation time and is considered as future work.

(S. Olafsson *et al*, 2006) shows the operation and research in data mining. The existing contributions of optimization methods in data mining touch on almost every part of the data mining process, from data visualization and pre-processing, to inductive learning, and selecting the best model after learning. It also shows applications related to the area of management of electronic services, namely customer relationship management and personalization and the OR community has over the past several years made highly significant contributions to the growing field of data mining. It concludes. Interest in data mining continues to grow in both academia and industry and most data mining issues where there is the potential to use optimization methods still require significantly more research.

(Binitha S *et al*, 2012) shows the different Bio-Inspired Algorithm used in optimization. applications and growth of natural computing in the last years is very drastic and has been applied to numerous optimization problems in computer networks, control systems, bioinformatics, data mining, game theory, music, biometrics, power systems, image processing, industry and engineering, parallel and distributed computing, robotics, economics and finance, forecasting problems, applications involving the security of information systems etc. And also conclude that there still remain significantly challenging tasks for the research community to address for the realization of many existing and most of the emerging areas in technology.

(M.A. Nada 2009) hybridized the Ant Colony algorithm and Genetic programming algorithm to integrate the movements of ant towards solution. Ants deposit a certain amount of pheromone depend on the quality of the solution found. Subsequent ants use the pheromone information as a guide toward promising regions of the search space. The genetic programming paradigm permits the evolution of computer programs which can perform alternative computations conditioned on the outcome of intermediate calculations.

(Y.Wang *et al*, 2013) integrates the DM and OLAP technologies. OLAP and data mining technology are both strengths, but also have shortcomings. With the combination of OLAP data cubes and data warehouse technology, a new data mining technology will better suit the actual needs. In order to achieve enhanced when combined with OLAP efficiency and flexibility purposes and combines technology and association rule in mining algorithm together, and conduct an appropriate improvements cube at the same time. OLAP and DM technology integration also belongs to the loose integration, and not fully realize the seamless integration of OLAP and DM technology.

(H. Drias *et al*, 2010) proposed a large scale information retrieval aiming at contributing to web searching. The greater the number of documents belonging to the collection, the more powerful approach required. A Bees Swarm Optimization algorithm called BSO-IR is designed to explore the prohibitive number of documents to find the information needed by the user. Numerical results exhibit the superiority of BSO-IR on previous works in terms of scalability while yielding comparable quality and planned to plan to hybridize metaheuristic with the distributed information retrieval approaches.

(Kanimozhi *et al*, 2009) propose an approach to set suitable support thresholds for frequent itemset generation. To find the appropriate associations, the user has to specify support and confidence thresholds. Thresholds play major role in deciding the number of appropriate rules found. User has many problems in specifying the suitable thresholds, without the familiarity of itemsets and their frequency in the database. It proposes a confidence-lift-based support threshold which can be automatically set from the itemset support. Results showed that the proposed model is performing well and generates relevant rules without missing interesting rules.

(Liu *et al*, 1999) proposes a novel approach to solve the of rare item problem. Rare items problem is arise when some items appear very frequently in the data, while others rarely appear. If minsup is set too high, those rules that involve rare items will not be found. To find rules that involve both frequent and rare items, minsup has to be set very low. The proposal allows the user to specify multiple minimum supports to reflect the natures of the items and their varied frequencies in the database.

(Glover 1986) presents main four areas (1)controlled randomization, (2) learning strategies, (3) induced decomposition and (4) tabu search for artificial intelligence. Innovation of these areas in artificial intelligence is from integer programming.

(R. Agrawal *et al*, 1993) presents a algorithm for generation of significant association rule from large database. The algorithm incorporates buffer management and novel estimation and pruning techniques. Experimental results are taken from large database of retailing company. The estimation procedure exhibited high accuracy and the pruning techniques were able to prune out a very large fraction of itemsets without measuring them.

(R. Agrawal *et al*, 1994) proposes a novel algorithm for association from database of large transactions of sales. The presented algorithms were AprioriTid. These algorithms show better results from previous algorithms.

(C. Romero *et al*) explore the extraction of rare association rules when gathering student usage data from a Moodle system. . Rare-association rules are more difficult to mine using traditional data mining algorithms, since they do not usually consider class-imbalance and tend to be overwhelmed by the major class, leaving the minor class to be ignored. Enhanced Apriori Algorithm was used for better results.

(J. Vaidya *et al*, 2002) addresses the problem of association rule mining where transactions are distributed across sources. . Every site holds some attributes of each transaction, and the sites desire to collaborate to identify globally valid association rules. The major assistance of this paper are a privacy preserving association rule mining algorithm given a privacy preserving scalar product protocol, and an efficient protocol for computing scalar product while preserving privacy of the individual values.

The grand goal is to develop methods enabling any data mining that can be done at a single site to be done across various sources, while respecting their privacy policies.

Algorithms

A. Bees Swarm Optimization.

The bio-inspired approach BSO proposed in simulates the collective bees behavior. It is based on a swarm of artificial bees cooperating together to solve a problem. First, a bee named InitBee settles to find a solution presenting good features. From this first solution called Sref we determine a set of other points of the search space from where bees will undertake a profound and intensive search for other better solutions. This set of solutions called Search Area, is computed in such a way that the different points are very far from each other in order to cover a large part of the search space. Then, every bee will consider a solution from Search Area as its starting location in the search. After accomplishing its search, every bee communicates the best visited solution to all its congeners through a table named Dance. One of the solutions stored in this table will become the new reference solution during the next iteration. In order to avoid cycles, the reference solution is stored every time in a taboo list. The reference solution is chosen according to the quality criterion. However, if after a period of time, the swarm observes that the solution is not improved, it introduces a criterion of diversification preventing it from being trapped in a local optimum. The diversification criterion consists to select among the solutions stored in taboo list, the most distant one. The algorithm stops when the optimal solution is found or the maximum number of iterations is reached. The main framework of BSO is presented by Algorithm (Djenouri, Y *et al*, 2013).

Algorithm 1: BSO Algorithm:

(1) Sref← The solution found by InitBee.
(2) while i < Max-Iter and not stop do
(3) Insert Sref in taboo list.
(4) SearchArea(Sref).
(5) Assign a solution from SearchArea to each bee.
(6) for each bee k do
(7) Built-Search-Area(bee k).
(8) Store the result in the table Dance.
(9) end for
(10) Choose the new reference solution Sref.
(11) end while

B. Tabu Search

Tabu Search (TS) is first introduced by (Fred Glover, 1986) . Tabu Search, employs a different approach to doing exploration: it keeps around a history of recently considered candidate solutions (known as the tabu list) and refuses to return to those candidate solutions until they're sufficiently far in the past. The simplest approach to Tabu Search is to maintain a tabu list called L of some explored solutions. First, the initial solution is generated and its neighbors are explored. If the best neighbor belongs to L then this neighbor could not be considered in the next search iteration and it remains in the tabu list. This process must be repeated until the maximum number of iterations is reached or the given conditions are met. Algorithm 2 describes briefly the main steps of TS (Sean Luke, 2009).

Algorithm 2: TS algorithm.

(1) l ← Desired maximum tabu list length
(2) n ← number of tweaks desired to sample the gradient
(3) S ← some initial candidate solution
(4) Best ← S
(5) L ← {} a tabu list of maximum length l " Implemented as first in, first-out queue
(6) Enqueue S into L
(7) repeat
(8) if Length(L) > l then
(9) Remove oldest element from L
(10) R ← Tweak(Copy(S))
(11) for n−1 times do
(12) W ← Tweak(Copy(S))
(13) if W ∈/ L and (Quality(W) > Quality(R) or R ∈ L) then
(14) R ← W
(15) if R ∈/ L and Quality(R) > Quality(S) then
(16) S ← R
(17) Enqueue R into L
(18) if Quality(S) > Quality(Best) then
(19) Best ← S
(20) until Best is the ideal solution or we have run out of time
(21) return Best.

C. Apriori Algorithm

Apriori Algorithm used to mine the frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses *prior knowledge* of frequent itemset properties. Apriori employs an iterative approach known as a *level-wise* search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found. This set is denoted L_1. L_2 is used to find L_2, the frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_K requires one full scan of the database (Han,J *et al*, 2006).

Algorithm 3: Apriori Algorithm:
Method:
(1) L1 = find frequent 1-itemsets(D);
(2) for ($k = 2$; $L_{k-1} \neq \phi$; k++) {
(3) C_k = apriori_gen(L_{k-1});
(4) for each transaction $t \in D$ {
(5) C_t = subset(C_k, t);
(6) for each candidate $c \in C_t$
(7) c.count++;
(8) }
(9) L_K= { $c \in C_k$|c:count _ min sup}
(10) }

(11) return $L = \cup_K L_K$;

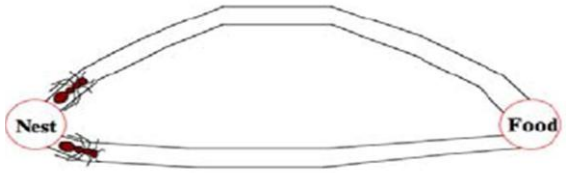
procedureapriori_gen (L_{k-1}:frequent (k-1)-itemsets):
(1) for each itemset $l_1 \in L_{k-1}$
(2) for each itemset $l_2 \in L_{k-1}$
(3) if ($l_1[1] = l_2[1]$) ^ ($l_1[2] = l_2[2]$) ^.....^ ($l_1[k-2] = l_2[k-2]$) ^ ($l_1[k-1] < l_2[k-1]$) then {
(4) $c = l_1 \bowtie l_2$
(5) if has_infrequent subset(c, L_{k-1}) then
(6) delete c;
(7) else add c to C_K;
(8) }
(9) return C_K;

procedure has infrequent subset(c: candidate k-itemset;
L_{k-1}: frequent (k-1)-itemsets);

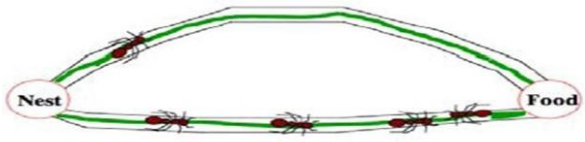
(1) for each (k-1)-subset s of c
(2) if $s \notin L_{k-1}$ then
(3) return TRUE;
(4) return FALSE;

D. Ant Colony Optimization

ACO was introduced in the early 1990s.It is based on the foraging behaviour of ants, which have the ability to select the shortest path among few possible paths connecting their nest to a food ground or site. A volatile chemical substance lay on the ground by the ants while walking and affecting in turn their moving decisions according to its local intensity is called pheromone which is the mediator of this behaviour (NN Das *et al*, 2013)



(a)



(b)

Figure 1: Principle of Ant Colony Optimization

Conclusion

Web contains noisy data, redundant information and which mirrored web pages in and abundance. Number of transaction occurred very frequently.In this survey we have studied mostly used technique that is apriori algorithm for finding the association between frequently occurred itemsets from large set of transactions. Apriori algorithm scans the dataset level wise which is purely based on prior knowledge and we have also surveyed that Bees Swarm Optimization and Ant Colony Optimization techniques helps to improve the quality of results whereas Tabu Search helped to increase the speed of optimization techniques to gain more in reasonable time.

References

Han,J., Kamber, J. and Pei, M. (2006), Data Mining: Concepts and Techniques, *Elsevier 2nd edition*.

KanimozhiSelvi C.S, and A. Tamilarasi (September 2009), An Automated association rule mining technique with cumulative support thresholds, *Int. J. Open Problems in Compt. Math*, 2(3), pp 427-438.

Liu, B., Hsu, W. and Ma, Y. (August 15-18, 1999), Mining association rules with multiple minimum supports, *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99), San Diego, CA, USA.*

Djenouri Y, Drias H., Habbas, Z., &Mosteghanemi H.(Dec. 2012), Bees Swarm Optimization for Web Association Rule Mining. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences*, 3, pp-142 – 146.

Djenouri Y., Amine Chemchem (Aug. 2013), A Hybrid Bees Swarm Optimization and Tabu Search Algorithm for Association, *Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress*, pp-120 -125.

Glover F.(1986), Future paths for integer programming and links to artificial intelligence, *Computers & Operations Research*, 13 (5), pp-533-549

Sean Luke, 2009, Essentials of Metaheuristics, available athttp://cs.gmu.edu/ !seanibookimetaheuristics/.

R. Agrawal, T. Imielinski and A. Swami (May 1993), Mining association rules between sets of items in large databases, *ACM SIGMOD Conference Washington DC, USA*.

R. Agrawal, R. Srikanth (1994), Fast algorithms for mining association rules, *VLDB Conference, Santiago, Chile*.

C. Romero, J.R. Romero,J.M. Luna,S.Ventura, Mining Rare Association Rules from e-Learning Data, *Dept. of Computer Science, University of Córdoba, Spain*, pp-171-180.

J. Vaidya, C. Clifton (2002), Privacy Preserving Association Rule Mining in Vertically Partitioned Data, *SIGKDD '02, Edmonton, Alberta, Canada, ACM*.

S. Olafsson, X. Li, S. Wu (2006), Operations research and data mining, *Elsevier*.

Binitha S, S Siva Sathya (May 2012), A Survey of Bio inspired Optimization Algorithms, *International Journal of Soft Computing and Engineering (IJSCE)*, 2(2), ISSN: 2231-2307.

Y.Wang, L. Yu (2013), Improved multi-level association rule in mining algorithm based on a multidimensional data cube, *IEEE*.

H.Drias, H. Mosteghanemi (2010), Bees Swarm Optimization based Approach for Web Information Retrieval, *International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM*, pp 6-13.

Nada M. A. Al Salami (2009), Ant Colony Optimization Algorithm,*UbiCC Journal*, 4(3), pp 823-826

NN Das, Anjali Saini (November 2013), A Study on Association Rule Mining Using ACO Algorithm for Generating Optimized ResultSet, *International Journal of Computer Science and Mobile Computing, IJCSMC*, 2(11), pp 123 –128.