International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-9 March-2013

Association Rules Optimization: A Survey

Anshuman Singh Sadh[1], Nitin Shukla[2]

Abstract

Association Rule mining is one of the important and most popular data mining technique. Association rule mining can be efficiently used in any decision making processor decision based rule generation. In data mining task in general we will find the frequent patterns to know the effective patterns from the huge data. Then we find positive and negative rules. If we observe the above phenomena then we come to the point that the rule generation is also huge. In this paper we survey several aspects of optimization techniques by which we can optimize the association rules. So the main motivation of our survey is to minimize the rule generation or optimize rule generation larger size of rules can be minimized.

Keywords

Association Rule Mining, Positive Association, Negative Association, Optimization.

1. Introduction

Data Mining is expected to relieve current mining methods from the sequential bottleneck, and provide the ability to scale to massive data sets and improve the response time [1]. Data mining has been a powerful technique in analyzing and utilizing data in today's information-rich society. However, privacy is nowadays a major concern in data mining applications, which has led to a new research area, privacy preserving data mining. A large amount of research work has been devoted to this area, and resulted in such techniques as k-anonymity [2], data perturbation [3], [4], [5], [6], and data mining based on [7], [8].

Association Rule Mining (ARM) is one of the most used research are in data mining. ARM can be used for discovering hidden relationship between items. By given a user-specified threshold, also known as minimum support, the mining of association rules can discover the complete set of frequent patterns.

Anshuman Singh Sadh, M.Tech Scholar, SRIT, Jabalpur (M.P).
Nitin Shukla, Assistant Professor, SRIT, Jabalpur (M.P).

That is, once the minimum support is given, the complete set of frequent patterns is determined [9]. In order to retrieve more correlations among items, users may specify a relatively lower minimum support[9]. Such a lower support often generates a huge amount of frequent patterns; but most of the patterns are already known or not interested to users. It is a tedious task for users to filter out these valueless patterns.

ARM is also studied in terms of market basket analysis, which is the analysis of the itemset which can be analyzed after the customer purchasing in the mall [9]. It is just like the analysis of the customer of purchasing behavior. Association rules also used in various areas such as telecommunication networks, market, risk management and inventory control etc. [9][10].

In [11] author suggests that Data mining [1] is used everywhere and large amounts of information are gathered: in business, to analyses client behavior or optimize production and sales [2].This signifies the research direction in several fields. We can use ARM and data mining application in health care, medical database, classification and combining these techniques with other approach extensively increases the potential behavior and applicability.

Our papers main motivation to survey in the direction of rule generation that is positive and negative association rule. We also emphasize to optimize the association rule so that it saves the time and un-wanted rules can be avoided.

In [12] author suggests that many of the researchers are generally focused on finding the positive rules only but they not find the negative association rules. But it is also important in analysis of intelligent data. It works in the opposite manner of positive rule finding. But problem with the negative association rule is it uses large space and can take more time to generate the rules as compare to the traditional mining association rule [12]. So better optimization technique can find a better solution in the above direction.

We provide here a brief survey on ARM. Other sections are arranged in the following manner:

Section 2 introduces negative and positive association rules; Section 3 discusses about optimization techniques; Section 4 describes about Literature Review; section 5 shows the analysis; Section 6 describes Conclusion.

2. Rule Generation

Association rule mining can be represent in terms of A ⇒ B [S, C] where A and B are sets of items; S is the support of the rules, defined as the rate of the transactions containing all items in A and all items in B i.e. Support (A ⇒ B) = P (A ∪ B) and C is the confidence of the rule, defined as the ratio of S with the rate of transactions containing A i.e. P (B / A). Support and confidence are measures of the interestingness of the rule. We calculate the support value for justifying the usefulness of the items present in the data set. A Higher support value indicates the effectiveness for the enterprise. The confidence signifies the decision theory, higher the confidence higher the decision accuracy.

In the application domain, items correspond to web resources, while transactions correspond to user sessions. So the equality of rule signifies that the domain is same and it can acquire the same alike set in the frequent pattern generation. The basic algorithm called Apriori Algorithm for finding the association rules was proposed and later modified uses Breadth First Search, Bottom Up Approach and performs well when the Frequent Items are short and and the use is easy. In this term we also discuss the term candidate generation like in the apriori algorithm, where it can generate 1-itemset,2-itemset and n itemset according to the data set present.

As an example of an association rule, we can think of the case of a super market. An association rule might, then, be in the following form: 'If a customer buys pork steaks, he buys at least two bottles of Coke as well'. In other words, it is in the form X => Y, where X and Y are items or set of items from the super market's database. In the case of City University, an example of an association rule might be that if a student wishes to do business studies, then with a probability of 90% chooses City.

The problem of identifying association rules was first introduced in (Agrawal, 1993). In (Hipp and Guntzer and Nakhaeizadeh, 2000) the formal description of the problem is given as follows: "Let ;={x1,...,xn} be a set of distinct literals, called items. A set X⊆X; with k=|X| is called a k-itemset or simply an itemset. Let a database D be a multi-set of subsets. Each T∈ D is called a transaction. We say that a transaction T∈ D supports an itemset X⊆X ; if X ⊆T holds. An association rule is an expression X =>Y, where X,Y are itemsets and X ∩Y = Ø holds. The fraction of transactions T supporting an itemset X with respect to database ' is called the support of X, supp(X)= |{T∈D | X ⊆T}| / |D|. The support of a rule X =>Y is defined as supp(X =>Y) = supp(X UY). The confidence of this rule is defined as conf(X =>Y) = supp(X U Y)/ supp(X)". The latter implies that we are looking at the fraction of transactions that contain the X itemset to see how many contain the Y itemset as well.

3. Optimization Techniques

There are several optimization techniques which can be apply on the association rule mining. Some of techniques are as follows:

Ant Colony Optimization
The Ant Colony Optimization algorithm is mainly inspired by the experiments run by Goss et al. [13] which using a grouping of real ants in the real environment. They study and observe the behavior of those real ants and suggest that the real ants were able to select the shortest path between their nest and food resource, in the existence of alternate paths between the two. This ant behavior was first formulated and arranged as Ant System (AS) by Dorigo et al. [14][15]. Based on the AS algorithm, the Ant Colony Optimization (ACO) algorithm was proposed [16]. In ACO algorithm, the optimization problem can be expressed as a formulated graph G = (C; L), where C is the set of components of the problem, and L is the set of possible connections or transitions among the elements of C.

Particle Swarm Optimization
PSO is a global optimization search algorithm introduced by Kennedy and Eberhart. It is an evolutionary computation technique discovered through simulation of collective social behavior such as bird flocking and fish schooling [17]. In PSO, particles represent candidate solutions in a solution space, and the optimal solution is found through moving the particles in the solution space. Individual particle flies through S-dimensional search space with velocity dynamically adjusted according to its own flying experience and its group's flying experience. The velocity and position of the particles are adjusted according to the following equation:
$V_{id(t)}=W.V_{id(t-1}+C_1.R_1.(P_{id(t-1)}-X_{id(t-1)})$ +……+….

$X_{id(t)}=X_{id(t-1)+}$ Vid(t-1)
Where d represents the value 1,2,3…. t represents the looping; Vi and Xi are the velocity and position of the ith particle; Pi is the previous personal best position of particle i and is called pbest; Pg is the previous best position of all the particles and is called gbest; W is the inertia weight; C1 and C2 are positive constants, and R1, R2.

Genetic Algorithm
Genetic algorithms work with a population of the potential solutions[18]. In computing terms, genetic algorithms map strings of numbers to each potential solution. Each solution becomes an individual in the population, and each string becomes a representation of an individual [18]. There should be a way to derive each individual from its string representation. The genetic algorithm then manipulates the most promising strings in its search for an improved solution. This algorithm follows the following cycle.
1. Creation of a population of strings.
2. Evaluation of each string.
3. Selection of the best strings.
4. Genetic manipulation to create a new population of strings.

4. Literature Review

In 2010 Ashutosh Dubey et al. [19] proposed a novel data mining algorithm named J2ME-based Mobile Progressive Pattern Mine (J2MPP-Mine) for effective mobile computing. In J2MPP-Mine, they first propose a subset finder strategy named Subset-Finder (S-Finder) to find the possible subsets for prune. Then, they propose a Subset pruner algorithm (SBPruner) for determining the frequent pattern. Furthermore, they proposed the novel prediction strategy to determine the superset and remove the subset which generates a less number of sets due to different filtering pruning strategy. Finally, through the simulation their proposed methods were shown to deliver excellent performance in terms of efficiency, accuracy and applicability under various system conditions. Means if optimization is achieved the data mining is also easily supported on mobile devices.

In 2012, Nikhil Jain et al. [12] discuss about Association rule mining. They suggest that association rule play important rule in market data analysis and also in medical diagnosis of correlated problem. For the generation of association rule mining various technique are used such as Apriori algorithm, FP-growth and tree based algorithm. Some algorithms are wonder performance but generate negative association rule and also suffered from Superiority measure problem. They proposed a multi-objective association rule mining based on genetic algorithm and Euclidean distance formula. In this method we find the near distance of rule set using Euclidean distance formula and generate two class higher class and lower class .the validate of class check by distance weight vector.

In 2012, Ashutosh Dubey et al. [20] Proposes an efficient method for knowledge discovery which is based on subset and superset approach. In this approach they also use dynamic minimum support so that we reduce the execution time. A frequent superset means it contains more transactions then the minimum support. It utilize the concept that if the item set is not frequent but the superset may be frequent which is consider for the further data mining task. By this approach we can also find improved association, which shows that which item set is most acceptable association with others. A frequent subset means it contains less transactions then the minimum support. It utilizes the behavior that the less count may be frequent if we attached the less count with the higher order set. Here we also provide the flexibility to find multiple minimum supports which is useful for comparison with associated items and dynamic support range. Their algorithm provides the flexibility for improved association and dynamic support. Comparative result shows the effectiveness of their algorithm.

In 2012, Preeti Khare et al. [21] discusses that importance of data mining is increasing exponentially since last decade and in recent time where there is very tough competition in the market where the quality of information and information on time play a very crucial role in decision making of policy has attracted a great deal of attention in the information industry and in society as a whole. They use density minimum support so that they reduce the execution time. A frequent superset means it contains more transactions then the minimum support. It utilize the concept that if the item set is not frequent but the superset may be frequent which is consider for the further data mining task. By this approach they can store the transaction on the daily basis, then they provide three different density zone based on the transaction and minimum support which is low(L), Medium(M),High(H). Based on this approach they categorize the item set for pruning. Their approach is based on apriori algorithm but provides better reduction in time because of the prior separation in

the data, which is useful for selecting according to the density wise distribution in India.

In 2012, Leena A Deshpande et al. [22] discusses about Semi-structured data which are a huge amount of complex and heterogeneous data sets. Such models capture data that are not intentionally structured, but are structured heterogeneously. These databases evolve so quickly like run time report generated by ERPs, World-Wide Web with its HTML pages, text files, bibliographies, various logs generated etc. These huge and varied become difficult to retrieve relevant information User is often interested in integrating various formats (like in biomedical data text, image or structured) that are generally realized as files, and also wants to access them in an integrated fashion. Users not only query the data to find a particular piece of information, but he is also keen in knowing better understanding of the query. Because of this variety, semi-structured DBs do not come with a conceptual schema.

In 2012, Smruti Rekha Das et al. [23] discusses about Support vector machine (SVM) which has become an increasingly popular tool for machine learning tasks involving classification, regression or novelty detection. SVM is able to calculate the maximum margin (separating hyper-plane) between data with and without the outcome of interest if they are linearly separable. To improve the generalization performance of SVM classifier optimization technique is used. According to the authors Optimization refers to the selection of a best element from some set of available alternatives. Particle swarm optimization (PSO) is a population based stochastic optimization technique where the potential solutions, called particles, fly through the problem space by following the current optimum particles. They used Principal Component Analysis (PCA) for reducing features of breast cancer, lung cancer and heart disease data sets and an empirical comparison of kernel selection using PSO for SVM is used to achieve better performance. This paper focused on SVM trained using linear, polynomial and radial basis function (RBF) kernels and applying PSO to each kernels for each data set to get better accuracy.

In 2012, Sanat Jain et al. [24] present an Apriori-based algorithm that is able to find all valid positive and negative association rules in a support confidence framework. The algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. The complexity and large size of rules generated after mining have motivated researchers and practitioners to optimize the rule, for analysis purpose. Their optimization done using Genetic Algorithm.

5. Analysis

After analyzing several research works in this direction, we come with following analysis:

1) Optimize association rules are also useful in the case of mobile data mining.
2) We can apply ACO and PSO for rule optimization.
3) Genetic algorithm is already applied in case of rule optimization [24].
4) Data partitioning is also needed, so that it can maintain the partitioning speed.
5) Subset and superset based distribution is also useful in the case of rule reduction.

6. Conclusion and Future Suggestions

In this paper we survey about association rule mining, negative and positive rule generation and optimization. We discuss different optimization methods. We also survey related research in the direction of rule optimization and provide the analysis. Based on this we suggest some further suggestions which are as follows:

1) Rule optimization can be applied using ACO and PSO and compare with the genetic algorithm.
2) After applying rule optimization, we can obtain reduced rules which are helpful in the case of handheld device.
3) Rule optimization is also done after partitioning.

References

[1] J.Z.Mohammed.Parallel and distributed data mining: an introduction. M.J.Zaki, C.-T.Ho (Eds.) Large-scale parallel data mining, Lecture Notes in Artificial Intelligence, 1759:1-23, 2000.
[2] Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557-570, 2002.
[3] Agrawal, R. and Srikand R. Privacy preserving data mining. In Proc. Of ACM SIGMOD Conference, pp. 439-450, 2000.
[4] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp. 589-592, 2005.

[5] Xu, S., Zhang, J., Han, D. and Wang J. Singular value decomposition based data distortion strategy for privacy distortion. Knowledge and Information System, 10(3):383-397, 2006.
[6] Mukherjeee, S., Chen, Z. and Gangopadhyay, A. A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier related transforms. Journal of VLDB, 15(4):293-315, 2006.
[7] Vaidya, J. and Clifton, C. Privacy preserving k-means clustering over vertically partitioned data. In Prof. of ACM SIGKDD Conference, pp.206-215, 2003.
[8] Vaidya, J., Yu, H. and Jiang, X. Privacy preserving SVM classification. Knowledge and Information Systems, 14:161-178, 2007.
[9] Ms. Kumudbala Saxena, Dr. C.S. Satsangi, "A Non Candidate Subset-Superset Dynamic Minimum Support Approach for sequential pattern Mining", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-4, Issue-6, December-2012.
[10] Manish Shrivastava, Kapil Sharma, Angad Singh, "Web Log Mining using Improved Version of Proposed Algorithm", International Journal of Advanced Computer Research (IJACR),Volume 1,Number 2,December 2011.
[11] Pragati Shrivastava, Hitesh Gupta," A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.
[12] Nikhil Jain,Vishal Sharma,Mahesh Malviya, "Reduction of Negative and Positive Association Rule Mining and Maintain Superiority of Rule Using Modified Genetic Algorithm", International Journal of Advanced Computer Research (IJACR) ,Volume-2, Number-4, Issue-6, December-2012.
[13] Goss, Simon, Serge Aron, Jean-Louis Deneubourg, and Jacques Marie Pasteels. "Self-organized shortcuts in the Argentine ant." Naturwissenschaften 76, no. 12 (1989): 579-581.
[14] M. Dorigo, Gianni Di Caro, and Luca M. Gambardella. Ant Algorithms for Discrete Optimization. Technical Report Tech. Rep. IRIDIA/98-10, IRIDIA, Universite Libre de Bruxelles, Brussels, Belgium, 1998.
[15] M. Dorigo and M. Maniezzo and A. Colorni. The Ant Systems: An Autocatalytic Optimizing Process. Revised 91-016, Dept. of Electronica, Milan Polytechnic, 1991.
[16] M. Dorigo and G. Di Caro. New Ideas in Optimisation. McGraw Hill, London, UK, 1999.
[17] J. Kennedy, R. Eberhart . Particle swarm optimization. In Proceedings of IEEE international conference on neural networks,Piscataway, NJ, USA (pp. 1942–1948),1995.
[18] Priyanka Gonnade, Sonali Bodkhe, "An Efficient load balancing using Genetic algorithm in Hierarchical structured distributed system", International Journal of Advanced Computer Research (IJACR),Volume-2, Number-4, Issue-6, December-2012.
[19] Ashutosh K. Dubey and Shishir K. Shandilya," A Novel J2ME Service for Mining Incremental Patterns in Mobile Computing",Communications in Computer and Information Science, 2010,Springer LNCS.
[20] Ashutosh Kumar Dubey, Animesh Kumar Dubey, Vipul Agarwal, Yogeshver Khandagre, "Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support",Conseg-2012.
[21] Preeti Khare, Hitesh Gupta, "Finding Frequent Pattern with Transaction and Occurrences based on Density Minimum Support Distribution", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.
[22] Leena A Deshpande, R.S. Prasad, "Efficient Frequent Pattern Mining Techniques of Semi Structured data: a Survey", International Journal of Advanced Computer Research (IJACR) Volume-3, Number-1, Issue-8, March-2013.
[23] Smruti Rekha Das, Pradeepta Kumar Panigrahi, Kaberi Das and Debahuti Mishra, "Improving RBF Kernel Function of Support Vector Machine using Particle Swarm Optimization", International Journal of Advanced Computer Research (IJACR) Volume-2, Number-4, Issue-7, December-2012.
[24] Sanat Jain, Swati Kabra, "Mining & Optimization of Association Rules Using Effective Algorithm", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4, April 2012.