International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 15484-15488

A Review on Association Rule Mining Techniques with Respect to their Privacy Preserving Capabilities

[1]M.Supriyamenon and [2]Dr. P.Rajarajeswari

[1]Research Scholar, [2]Professor

[1,2,] Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation (KLEF),
Guntur, Andhra Pradesh India.

[1]Orcid: 0000-0001-6651-2218

Abstract

Datamining, an urging requirement in the current era and whose scope of research is expected to be for upcoming decades. Among the well versed techniques of data mining association rule mining plays a prodigious role. This technique emphasizes on curious association, correlations, frequent patterns etc. from the given data sources to be mined. The primary task of association mining resides in uncovering the frequent patterns and exploring the association rules. Multiple variation of association rule mining algorithms with regard to their performance factors are available. One important constraint entailing the extraction of association rules is, privacy preserving of sensitive data. There is a need to maintain a sustainable ratio between protection of privacy and knowledge discovery. This paper enlightens and reviews the various association mining techniques with respect to their privacy preserving capabilities and also adds a pinch on the tools preferred for such privacy.

Keywords: -ECLAT, FP, GROWTH, AIS, PS.

INTRODUCTION

The art of discovering unseen and useful information from large heterogeneous data sources in a skilled manner is called data mining. Data mining discloses knowledge out of knowledge base or data warehouses. Knowledge is something which is of interest to the user. Knowledge helps in better decision making. Therefore extracting such knowledge is very vital. Data mining task includes several techniques like classification, clustering, association rule mining, regression etc. Amongst them association rule mining is one of utmost interest. Different techniques for association rule mining with respect to performance factors are considered and a comparative analysis is reviewed. This review also considers some of the privacy factors which well suit their respective association rule mining technique. Association rule mining algorithm help in extracting desired association rules by considering the given data set. They are intended to recognize strong rules basing on the measures of in testiness. Association rule mining is given in terms of support and confidence where strong rulesbasing on their interestingness. Association rule mining is given in terms of support and confidence where they are filter ie biased if they don't satisfy value considered.

$$\text{Support (xy)} = \frac{\text{Transactional support (xy)}}{\text{Total no. of Transactions in D}} \quad (1)$$

$$\text{Confidence (xy)} = \frac{\text{Support (xy)}}{\text{Support(x)}} \quad (2)$$

Support reflects frequency of the item and confidence reflects the no of transactions containing if/then pattern. An association rule satisfying a predefined threshold (breakeven) value for support and confidence are only considered for rule exploration. We focus on many such techniques for disclosing strong association rules in our work. Remaining part of this paper is organized as below: Section 2 discuss the related work. Section 3 elaborates the methodology and reviews various association rule mining algorithms in various aspects. Section 4 discuss about privacy preserving of association rules using various approaches. Section 5 presents a related mapping between association algorithms, rules and privacy approaches. Finally conclusion based on above features is presented in section 6.

RELATED LITERATURE WORK/ BACKGROUND

Kanithakelkar, smithaR.sankher[11] discussed briefly on the optimization factors like execution time for legacy apriori and its improved version(direct find and remove DFR) which targeted at improving the efficiency of association rules. This method eliminates those subsets which are not frequent their by reducing the redundant generation of sub item sets during preserving candidate itemsets.Increased probability for information in scanning databases and reduced size of itemsetsamitmittal, karthikguptaetc [18] analyzed frequent item set mining in transactional database. It aims at reducing number of scan and henceforth improving efficiency.

T. Bharathi , P.Krishnakumari provided comparative chart for mapping various association rule mining algorithms with their

accuracy rates ,application, speeds etc. this given is a scope for detailed analysis.

S. Patel Tushar, mayurpanchaletc [15] reviewed few unique factors and inside working various mining algorithms i.e. each and every strengthening factors and cons of algorithms like apriori , DHP, partitioning, Eclat , FP Growth to their fullest. Author proved that merge and split (SaM) is better algorithm than others.

METHODOLOGY OF VARIOUS ASSOCIATION RULE MINING ALGORITMS

Apriori Algorithm

This is an algorithm given by R.Agarwal for discovering frequent Item sets , based on the principal /property that all sub sets of frequent item sets must be frequent .[13] This algorithm extends frequent sub sets one at a time to bring out generation of candidate .Such candidate groups are tested with data. This algorithm works in 2 phases join and prune. Apriorisearch resembles a BFS which use hash tree Data structure for counting candidate item sets.

Apriori TID

AprioriTIDalsoaims at disclosing frequent item sets in a transaction database. It is an alternate version of Apriori which uses Apriori to find candidate item sets before beginning of pass.Anintresting part of AprioriTID is that the data base is not considered for support counting after first pass. Apriori TID always scans candidate set for support counts . Not preferable when size of problem grows[9].

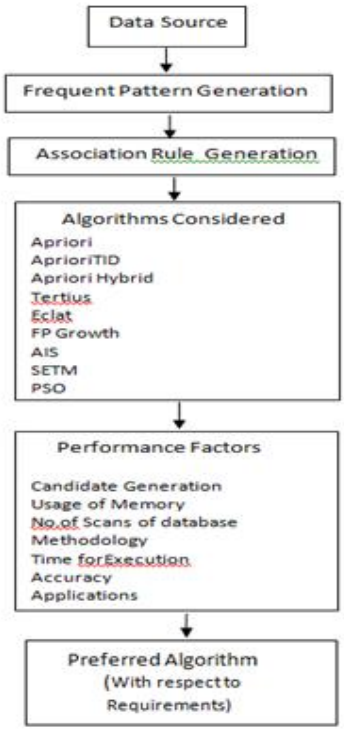

Figure 1: Comparative study is shown in the form of system Architecture

AprioriHybrid

This algorithm is a combination of Apriori and AprioriTID. It uses the basic Aprioriin the initial pass but Apriori TID glitters in the next passes[10]. This gives better results in many cases.This algorithm includes extra cost when sliding from Apriori to AprioriTID .

Tertius Algorithm

It walks with order logic representation. It employs rules according to the confirmation measures. An inclusion of several options such as class index, frequency, classification, confirmation threshold, Horn clauses, missing values ,noise threshold , roc analysis etc are with Tertius. It suffers from heavy runtime which is based on the literal count in the rules. An increment in Literals indulges an exponential raise (max is 3 preferable) but still takes hours.[13]

Eclat

Eclat algorithm projects transaction as bit matrix and rows projected the item sets support. It follows a depth first transversal of prefix tree.[11]

Bit matrices

Here rows present items and columns represent transactions. If a respective item is seen in a transaction, its bit is set else cleared.

Search tree transversal

Eclat searches the prefix tree using depth first order. A modification of node to its child is done by using a new bit matrix which is an intersection of first row with all rows followed. Similarly for other children also rows with infrequent item sets are deleted from matrix.

FP Growth

The frequent pattern tree (FP-Tree) is a structure that's compact and stores quantitative information about frequent pattern in data base. It applies divide and conquer strategy.[12] It first compresses the input db creating an FP tree instance to represent frequent item .Next it divides the compressed db into a set of condition databases. Each one associated with one frequent pattern. Finally each such db is mined separately. This way ,FP growth reduces the search cost looking for short patterns recursively and then concatenating them in long frequent patterns, offering best

selectivity. For large dbs, FP tree cannot be held in main memory, so cure is simple partition into smaller database and then construct FP tree.

AIS

This algorithm is a gift by Agarwal[15], Imielinske and Swami. It improves the face of databases for better decision making. It generates only one item consequent association rule that is only rules such as A∩B=>C Not A =>C∩B . It runs in 2 phases that is frequent item set generation and then followed by exploring confident and Frequent association rules. It needs the data set to be scanned several number of times for frequent item set and then rules. During first pass individual items support count is considered and then based on threshold value minimum supported item sets are removed/washed out.

ASSOCIATION RULE HIDING TECHNIQUES

Preserving privacy in mining investigates the backlogs of data mining methods that pop out from penetration into privacy issues of individuals and organizations. Some of such privacy preserving techniques are classified into broad categories like Heuristic based, Border based, Exact based, Reconstruction based and cryptographic based approaches.

Heuristic Based

This approach targets at taking a best decision that suits hereby considering maintaining secrecy of sensitive knowledge. It comes in two flavorsie hiding either by *Distortion*, meaning changing old with new. [5] But some applications may not allow such privacy as modified may not be acceptable. So Data *blocking* may be preferred where instead of changing sensitive data is blocked.

Border based approach

This approach hides data only at borders ie by constructing a border between frequent and infrequent item sets [6]. Thereafter this method selects minimum side for modifications which aim at privacy and also targets at high quality result.

Exact approach

This method is a non-heuristic one which is based on constraint ie one which satisfies constraint is selected and binary integer programming, method is applied to solve. This approach inculcates no side effects. This gives best solution comparatively. [16] This shortens the distance between the original one and purified one.

Reconstruction based approach

This is one designed for designing privacy engaged databases by focusing on sensitive items of source database. Analyzed by Mielikainen and proved that the side effects of proposed one are small compared to heuristic. Guo[17] also presented his work which stands on pillars of FP-Tree which is capable of reconstructing original one from non-characterized data thereby generating secure repositories.

Cryptographic based approach

This method is preferable when several parties are involved and their data are spread over geographically i.e. distributed. This includes communication and sharing between parties yet maintaining their boundaries. This method paves in 2 ways Horizontal partition distributed data and vertically partitioned distributed data. [3]Vaidya and Clifton gave their finding and outcomes for vertical ones. Finally we present a tabular representation of our survey and also add a flavor of Inferences from the various mining techniques and their paired privacy preserving preferences.

Table 1: Analysis 1

SNO	PROPERTIES	APRIORI	APRIORITID	APRIORIHYBRID	TERTIUS
1	CANDIDATE GENERATION	Apriori produces candidate item sets from of previous pass by not taking the transaction in database	Once the first pass is completed, database is not considered for counting support of candidate itemsets	It generates Candidate item sets by using Apriori but later jumps to AprioriTID	Candidates are generated by considering attribute pairs for the rule generation.
2	METHODOLOGY	Join and prune phases/steps	Considers Join and Prune in combination with TIDS	Combination ofApriori and AprioriTID	First order logic presentation is preferred
3	DATABASE SCAN	Needs many scans of databases	Needs only one scan	Addition of Apriori and AprioriTID	Scan depends on the count of literals in rules
4	MEMORY USAGE	It occupies high memory space for the process of candidate generation	In first pass, this algorithm needs memory for Lk-1 and Ci-1 candidate generation. It indulges extra cost in case if it does not fit in memory.	It infers extra memory when sliding from Apriori to AprioriTID	Consumes considerable time and printsout rules when program runs short of memory and messages
5	EXECUTION TIME	Mainly spends more for Candidate Generation	Executes fast in contrast to Apriori for small problems but incurs more time for large ones.	Preferably better than Apriori and AprioriTID	Consumes considerable long time for larger sets ie even hours
6	DATA SUPPORT	Limited	Nearly large sets	Very Large datasets	Limited
7	ACCURACY	Less	Better than Apriori	Increased Accuracy compared to AprioriTID	Considerable not high ie Average
8	Applications	Can be mainly preferred for closed Item sets.	Preferred for small problems.	Well suited for closed sets.	Most generally preferred.
9	PRIVACY PRESERVING APPROACH PREFFERED	Heuristic approach	Exact approach	Heuristic approach	Cryptographic approach

Table 2: Analysis 2

SNO	PROPERTIES	ECLAT	FP GROWTH	AIS	PSO
1	CANDIDATE GENERATION	Uses bit matrix representation of transactions &prefix tree in dfs order.	Does not generate candidate set but takes few pases over database	Scans database each time for generating candidates.	Candidates generated when db scan in progress
2	METHODOLOGY	Bit matrix rep.& depth first search of prefix tree construction from bit matrix	Two phases of divide and rule method	2 stages,1ST frequent item set generation	Uses concept of ‘neighbourhood’
3	DATABASE SCAN	Only once till matrix construction	Scans fewer no of times until construction of FP-tree.	Multiple Scans	Same asAIS
4	MEMORY USAGE	Considerably less as prefix tree rep. is considered.	Comparitively Average	Occupies much space	Less Comparitively
5	EXECUTION TIME	Faster Initially and average later	Average	Long Time	Very Fast
6	DATA SUPPORT	Large	Very Large	Less	Very Large
7	ACCURACY	Considerably better	High when compared with AprioriTID	Too small or less	Excellent
8	APPLICATIONS	Mainly preferred for free Itemsets.	Preferred for large Applications	Well suited for small problems	Large scale including closed sets and free item sets etc.
9	PRIVACY PRESERVING APPROACH PREFFERED	Reconstruction based approach	Reconstruction based approach	Exact approach	Cryptographic approach

CONCLUSION

Our paper reviewed and had a comparative analysis of various Association Rule mining algorithms with regard to various characteristics .On the basis of runtime and theoretical considerations available algorithms are systemized and performances are analyzed by even proposing the best privacy preserving technique that suits the algorithms. We conclude that apriori algorithm's cons are overcome by Apriori & AprioriTID. For closed Item sets Apriori lead sand for free Eclat wins. For larger applications PSO makes it.

REFERENCES

[1] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco, USA, 2001, ISBN 1558604898.

[2] R.Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules," In Proc. of VLDB '94, pp. 487-499, Santiago, Chile, Sept. 1994.

[3] J. Vaidya& C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," In proc. Conf. Knowledge Discovery and Data Mining, pp. 639–644, July 2002.

[4] Komal shah, Amitthakkar & Amitganatra, "Association Rule Hiding by Heuristic Approach to Reduce Side Effects & Hide Multiple R.H.S. Items" International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012.

[5] Elham Hatefi, Abdolreza Mirzaei & Mehran Safayani, "Privacy Preserving Quantitative Association Rule Mining Using Convex Optimization Technique", 2014 7th International Symposium on Telecommunications (IST'2014).

[6] Tapan Sirole & Jaytrilok Choudhary, "A Survey of Various Methodologies for Hiding Sensitive Association Rules" International Journal of Computer Applications (0975 - 8887) Volume 96-No.18, June 2014.

[7] MinubhaiChaudari,JigarVarmora,"Advance Privacy Preserving in Association rule Mining",in IEEE conference (ICEEOT) 2016 .

[8] Tarinder Singh; ManojSethi, "Sandwich-Apriori: A combine approach of Apriori and Reverse-Apriori",2015 Annual IEEE India Conference (INDICON) Year: 2015 Pages: 1 - 4.

[9] Zhi-Chao Li; Pi-Lian He; Ming Lei, "A high efficient AprioriTid algorithm for mining association rule", 2005 International Conference on Machine Learning and Cybernetics, Year: 2005, Volume: 3,Pages: 1812 – 1815.

[10] Surendra Kumar Chadokar; Divakar Singh; Anju Singh, "Optimizing network traffic by generating association rules using hybrid apriori-genetic algorithm",2013 Tenth International Conference on Wireless and Optical Communications Networks (WOCN),Year: 2013,Pages 1 – 5

[11] Manjitkaur, UrvashiGrag, "ECLAT Algorithm for Frequent Itemsets Generation", International Journal of Computer Systems (ISSN: 2394-1065), Volume 01– Issue 03, December, 2014.

[12] Li Min; Wang Chunyan; Yan Yuguang,"The Research of FPGrowth Method Based on Apriori Algorithm in MDSS", Digital Manufacturing and Automation (ICDMA), 2010 International Conference on Year: 2010, Volume: 2 Pages: 770 - 773.

[13] Sunita B. Aher, Lobo L.M.R.J, "A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning " International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012.

[14] Radoslav Harman, "A very brief introduction to particle swarm optimization".

[15] Rakesh Agarwal, Tomasz Imielinski, Arun swami, "Mining association rules between stes of items in large databases", SIGMOD '93 Proceedings of /management of data.

[16] A. Gkoulaslas Divanis & V.S>Verykios, "An Inter programming Approach for Frequent Itemset Hiding, " (CIKM '06)

[17] Y.Guo, "Reconstruction-Based association Rule Hiding," IDAR-*Workshop on Innovative Database Research 2007.*