International Journal of Engineering &Technology, 7 (1.7) (2018) 59-61

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Frequent item set mining using normalized FP-growth algorithm

N. K. Manikandan [1]*, D. Manivannan [1]

[1] *Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai-62, TamilNadu, India*
**Corresponding author E-mail: manikandan1488@gmail.com*

Abstract

As the volume of data and its storage schemes are increasing drastically, the knowledge discovery from these huge volume of heterogeneous and high dimension data emerges as an essential process. Number of algorithms for data association analysis has been introduced considering time and main memory requirements. However this algorithms get completed when the items and records grows extremely high. In this paper we have created a data structure that can be reused without modifying the schema. So the aim of this work is to make an efficient association rule mining independent of the algorithm being selected.
Our data structure make data access much faster by simplifying and reorganizing them by implementing shuffling strategy using hamming distance and inverted index mapping. In this work we explore our algorithm's efficiency by using many datasets containing millions of records in it. We increased the runtime orders of the magnitude and reduced the auxiliary and main memory requirements.

Keywords: *Item Setmining; FP Growth Algorithm; Association Rule Mining.*

1. Introduction

As the data plays an efficient role knowledge discovery in all industries, we are in need of storing and maintaining the raw data to generate useful patterns with respect to the user's needs [1].
Data mining techniques are mainly used to retrieve useful information from a huge database. Also they mainly focused on prediction and description of given data. Number of techniques like classification, clustering, association rule mining are there to perform this tasks. Among these, Association Rule Mining (ARM) plays a vital role in discovering useful relationship among different attributes in huge database. ARM is used to create rules which obeys the minimum support and minimum confidence levels. During the process of association rule mining a set of frequent item sets are created as an initial step and then association rule mining are created depending on the frequent item sets.
Centralized and distributes are the two types of database environments exists. In distributed model, the data is partitioned into fragments and each fragment is assigned with a site. So obviously privacy issues will arise when sharing fragments across sites and the site owners may not be interested in sharing the private data to other sites but they will be interested in knowing the global results obtained from mining process.
In order to handle such a security of data in distributed systems emerged the privacy preserving data mining techniques. The main difference between the normal data mining techniques like classification, cluster, association and this privacy mining is that the normal algorithms mainly focus on extracting useful knowledge from the database whereas the latter is focused on handling sensitive data of the users' records [5].
The idea behind any privacy preserving data mining (PPDM) is to allow computation of aggregates on the whole data set without affecting the privacy of the data. All the site owners may be interested in collaborating to get the result but don't trust sites to share their own data sets.
More care should be taken on the privacy of data on the context of distributed mining of data. This can be made possible by implementing secured multi party computation technique. PPDM uses data mining techniques to implement data mining objectives without bothering the privacy of the data. Thus in many data mining applications privacy preserving techniques plays a major role.

2. Related works

2.1. Feature subset selection

Feature subset selection undergoes the process of selecting the subset of useful features from data that generate results similar to the original data set features. The resultant features would be most effective and efficient. The efficiency is reflected by its time taken to find a subset of a feature whereas the effectiveness is measured by the quality of the subsets of the feature.
Depending on this parameters a new algorithm called fast clustering-based feature selection algorithm (FAST) algorithm is introduced in this work. This algorithm works in two steps. In the initial step the data is divided into group using bipartite graph based clustering technique. Then in the second step, the features that are most relevant to the target classes are extracted from the cluster to create a subset of features.
Features of the clusters are independent and the strategy of clustering in FAST produce useful and independent features. To make FAST algorithm most efficient they implement minimum spanning tree (MST) based clustering method. They will be most effective and efficient when comparing to other feature selection algorithms like CFS, Consist, FCBF and FOCUS-SF. As a final result, on using the FAST algorithm in publicly available high

dimensional images, text data and microarray, it is shown that FAST is more accurate on four types of classifiers namely Naïve Bayes, C4.5, probability based and instance based IB1 techniques.
The embedded methods uses feature selection to train data and are concern with learning algorithms so that it is more efficient than other categories. Embedded approaches are utilized by machine learning algorithms like ANN and decision trees. The accuracy of the learning algorithms are used by wrapper methods to find the accuracy of the selected subsets. So the learning algorithms will be always accurate [16].
But the selected features will be less generic and much complex. And the filtering methods are not dependent on learning algorithms and it will be very generic. The computation is less complex but less accurate in its output. The feature subset selection is the action of removing the irrelevant and redundant features in the dataset. Because redundant features may forbid better prediction whereas the irrelevant features reduce the accuracy of the prediction. There are many feature section algorithms exists which very often just remove the redundant features and merely the irrelevant features. Very fewer techniques concentrate on both. FAST is such an algorithm which performs both the operations. A good example is Relief which can eliminate the irrelevant features from different targets based on distance based analysis. But it fails to remove redundant features because two features can be of similar weights. Relief-F is the extension of Relief used this method with incomplete and noisy data sets and deals with multi class problems but fails to handle redundancy elimination.

3. Association rule mining

The implementation of association rule mining for a large database of sales transaction has been introduced in this work which combines tow algorithms to form a single hybrid one. It address many existing problems in the association rule mining of datasets of huge volume and low dimensional data. On experimenting with the proposed Sensitivity Apriori Algorithm (SAA) which combines Tuple Apriori and Enhanced Apriori to discover association rule in large database of transactions. We compared this algorithms with the existing two algorithms namely ODARM and SETM and proved that it is more efficient than the above two. They are more efficient with respect to the problem size and they are more accurate in order of magnitude when applied with larger problems [11].
In order to resolve the issue of privacy preserving data mining for a shared confidential databases owned by two parties. They wish to share their data with each other without revealing the unnecessary information to each other. Our work is to handle this task by protecting the privileged information and to enable data for research.
We implement a decision tree learning with popular ID3 algorithm. We prove that our work is more efficient than any other generic solutions and needs very few rounds of communication and limited bandwidth.

3.1. Complications in frequency analysis

Here we assume an example of having a network which have three or four players processioning some private data which them and like to share some data with each other without revealing their own data to others [18]. They are permitted to share private messages with each other and to broadcast messages to the community as a whole. Many protocols are there in the market to implement this multiparty secure function evaluation task. But our new protocol is different from the computational complexity of the function which is computed from complexity of the protocol evaluates it.

4. Methodology

4.1. Data compression technique

Compression procedure is the first step after data is received. This procedure contains the work of data structure simplification, reducing the memory requirements and enabling speedy data access. Run Length Encoding is the compression technique used which arranges data with same values in a consecutive way.

1) The data size reduction is analysed with refer to the usage of data with ARM algorithms
2) To assess time based reduction ratio when data is used by Frequent Pattern (FP) Growth algorithms that detect the association rules that comply with certain constraints.
3) To assess the reduction in time of execution when ARM algorithms utilizes new data structures.

4.2. Enhanced frequent pattern algorithm (EFP)

Initially this algorithm investigates all the frequent patterns in the data and considers the item set which are in the frequency of the threshold mentioned.
In preprocessing, we implements splitting method to transform the database threats to individual privacy. Then differential privacy has been performed to address this problem. By using this differential privacy we can offer a strong guarantee of the privacy of the data without any wrong assumptions on the attackers' knowledge [13].
By adding some noise over the dataset we are making our data more insensitive to change in any individual records so that the privacy leak will be highly restricted. There are number of algorithms implemented for mining frequent datasets. The most popular algorithms are Apriori and FP growth. Apriori is meant for BFS (Breadth First Search) which uses candidate set generation and test algorithm. The FP growth algorithm is used to differentiate private FIM algorithm based on FP growth algorithm.
In this work, we learnt that the private FIM algorithm can not only attain high utility of data and high degree of privacy but also provide high time based efficiency. Our algorithm stands unique from any other FIM algorithm in satisfying all these requirements simultaneously. It is a big challenge to handle this privacy trade off. In our technique we considered two related settings. First we considered data owners and analyser as two different entities and the data is distributed to different parties who need to do data mining process.
After that the data records are secured from data miners and data owners will apply data perturbation so that any analysis of data trends can be performed without revealing the originality of records. Next we implement data mining which protects data records from access by other owners. We have created user view and admin view to view the user's details being processing using Apriori based association rule.
Algorithm EFP
Input: A heterogeneous database and user defined parameters for breadth first search m_i, L,p, t_m,t_l.
Output: dataset with feature extracted with maximal frequent itemsets
Initially table procedure is created for statistical table
Identify minsup for all the items to create association and to generate M_i.
Create set S_i for storing sensitive dataset

AEP :={{ e} | e ε M_i}

While (M_{k-1} ε E) do

Call the feature selection scheme to assess duplicate data.
Analysis the frequent and infrequent itemsets.
Find t_m the occurrence of repeated data by scanning each itemset
Find t_l the sensitivity of dataset by checking with primary key

Move repeated data from M_k to M_i
Move sensitive data to Si

Result: = Results U {M_i, S_i}

K=: k+1;

MFI = MFI U Results
Return MFI

5. Conclusion

The main issue in handling the privacy preservation of data in data mining is to analysis the features which are repeated and irrelevant. Both this data set will degrade the accuracy of the extraction and also produce useless results. In order to handle this two problem we proposed an algorithm which eliminate the irrelevant data by comparing the weightage of the dataset for its importance and the occurrence of itemsets in the database. By resolving this, we can use this database in any types of heterogeneous database.

References

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc 20th Int'l Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.

[2] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Conf., pp. 439-450, 2000. https://doi.org/10.1145/342009.335438.

[3] D. Beaver, S. Micali, and P. Rogaway, "The Round Complexity of Secure Protocols," Proc. 22nd Ann. ACM Symp. Theory of Computing (STOC), pp. 503-513, 1990. https://doi.org/10.1145/100216.100287.

[4] M. Bellare, R. Canetti, and H. Krawczyk, "Keying Hash Functions for Message Authentication," Proc. 16th Ann. Int'l Cryptology Conf. Advances in Cryptology (Crypto), pp. 1-15, 1996. https://doi.org/10.1007/3-540-68697-5_1.

[5] J.C. Benaloh, "Secret Sharing Homomorphisms: Keeping Shares of a Secret Secret," Proc. Advances in Cryptology (Crypto), pp. 251-260, 1986.

[6] J. Brickell and V. Shmatikov, "Privacy-Preserving Graph Algorithms in the Semi-Honest Model," Proc. 11th Int'l Conf. Theory and Application of Cryptology and Information Security (ASIACRYPT), pp. 236-252, 2005. https://doi.org/10.1007/11593447_13.

[7] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules," Proc. Fourth Int'l Conf. Parallel and Distributed Information Systems (PDIS), pp. 31-42, 1996. https://doi.org/10.1109/PDIS.1996.568665.

[8] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, "Efficient Mining of Association Rules in Distributed Databases," IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, Dec. 1996. https://doi.org/10.1109/69.553158.

[9] T. ElGamal, "A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms," IEEE Trans. Information Theory, vol. IT-31, no. 4, July 1985. https://doi.org/10.1007/3-540-39568-7_2.

[10] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 217-228, 2002. https://doi.org/10.1145/775047.775080.

[11] R. Fagin, M. Naor, and P. Winkler, "Comparing Information without Leaking It," Comm. ACM, vol. 39, pp. 77-85, 1996. https://doi.org/10.1145/229459.229469.

[12] M. Freedman, Y. Ishai, B. Pinkas, and O. Reingold, "Keyword Search and Oblivious Pseudorandom Functions," Proc. Second Int'l Conf. Theory of Cryptography (TCC), pp. 303-324, 2005. https://doi.org/10.1007/978-3-540-30576-7_17.

[13] M.J. Freedman, K. Nissim, and B. Pinkas, "Efficient Private Matching and Set Intersection," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), pp. 1-19, 2004. https://doi.org/10.1007/978-3-540-24676-3_1.

[14] O. Goldreich, S. Micali, and A. Wigderson, "How to Play Any Mental Game or a Completeness Theorem for Protocols with Honest Majority," Proc. 19th Ann. ACM Symp. Theory of Computing (STOC), pp. 218-229, 1987.

[15] H. Grosskreutz, B. Lemmen, and S. R€uping, "Secure Distributed Subgroup Discovery in Horizontally Partitioned Data," Trans. Data Privacy, vol. 4, no. 3, pp. 147-165, 2011.

[16] W. Jiang and C. Clifton, "A Secure Distributed Framework for Achieving k-Anonymity," The VLDB J., vol. 15, pp. 316-333, 2006. https://doi.org/10.1007/s00778-006-0008-z.

[17] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037,Sept. 2004. https://doi.org/10.1109/TKDE.2004.45.

[18] M. Kantarcioglu, R. Nix, and J. Vaidya, "An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining," Proc. 13th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 515-524, 2009. https://doi.org/10.1007/978-3-642-01307-2_48.

[19] L. Kissner and D.X. Song, "Privacy-Preserving Set Operations," Proc. 25th Ann. Int'l Cryptology Conf. (CRYPTO), pp. 241-257, 2005.