International Journal of Research and Advanced Development (IJRAD)

Knowledge Extraction of Infrequent Item Set using Association Rule Mining

Sadesh Selvaraj[1], R. Madhumeena[2], R. Madhumitha[3], K. Monica[4], V. Vinmathi[5]

[1]Assistant Professor, Department of Computer Science & Engineering,
Velalar College of Engineering and Technology, TN, India, sadesh_vcet@yahoo.com
[2, 3, 4, 5]IV Year B. E., CSE, Department of Computer Science & Engineering,
Velalar College of Engineering & Technology, Erode, TN, India.
rmadhumeena@gmail.com[2], madhumithagrk@gmail.com[3], monicasri17@gmail.com[4], vinmathivelu@gmail.com[5]

Received Date: 11th March, 2017, Revised Date: 29th March, 2017, Accepted Date: 14th April, 2017.

Abstract - The most common relation between different data is gathered using weighted item set mining. Frequent and infrequent item set are available in the dataset. Infrequent item set is the rarely used items in a database. Mining of frequent items is used for retrieving the most relevant data in the dataset. With the help of transactional dataset, as an input the weighting function is calculated. The minimum support value is used to calculate the infrequent item set support value. Then the addition of processes is done for all the systems separately. The two systems are then combined, and the minimal values are summated. Finally, the minimal among all the systems is considered. The threshold value is first calculated, and if the addition value is greater than the threshold value, then the combination of systems is not considered. Or else, it is considered for the next step. The equivalent weighted transaction dataset is then calculated from transaction dataset. Since the infrequent weighted item set minimum support value is known, it is easier to find the threshold value for the equivalent weighted data item set. And then the system summations are found out. Then an infrequent weighted itemset miner is used to find the relevant systems that present in the two results. By using required algorithms, we can find the infrequent weighted item set. And from that, the final result is calculated.

Keywords - Weighted item set mining, frequent sets, infrequent sets, threshold value, and transactional dataset.

I. INTRODUCTION

Now a day the knowledge discovery in all the industrial fields is in high demand. Hence it is important to store all the necessary data and to provide useful patterns for the user needs. The storage will be handled in a database maintained by the specific organizations. Many data mining techniques are currently in use to retrieve the relevant and useful information from the large set of databases. Data mining has two important goals such as prediction and description.

To attain these fundamental goals, there are many data mining techniques available such as association rules, classification, clustering and so on. Among these different techniques, association rule has many applications to find the relationship among different attributes in a large set of databases. Association rule mining helps in finding the rules which satisfy the user-specified minimum support and minimum confidence [13]. For finding association rules, the first step is to compute the frequent item set and then to generate the rules based on the frequent item set. Centralized and distributed are the database environments. The privacy conflict comes when the data is distributed among multiple sites, and no site owner is interested to expose their private information [6], but they are interested to know the results obtained from the mining process.

Privacy preserving is a new stream in data mining era which mainly focuses on incorporating the privacy in data mining techniques [2]. It is used to protect the confidential data of the user. There is an important difference between regular data mining algorithms under various data mining techniques such as classification, association, clustering. The privacy preserving data mining algorithms deals with analyzing the stored raw data and extracting the useful knowledge discovery patterns from the database [4].

The objective of many distributed methods for privacy preserving is to allow useful aggregate computations on the complete data set [5]. This is done by preserving the privacy of the individual sites data/information. Each site owner is interested to collaborate in obtaining combined results, but they do not completely depend on the other regarding the distribution of their data sets. The important property of any data mining system is privacy preserving of data/information [7]. Mainly in distributed data mining, privacy preserving is one of the most crucial aspects. Secure multi-party computation is a useful approach [15], and it makes use of a mining algorithm to procure data mining objectives without revealing private data [14].

II. RELATED WORKS

A. Fast Clustering-Based Feature Subset Selection - Algorithm for High-Dimensional Data

This technique aims to produce the relevant results as the original specifications [11]. This algorithm can be obtained from the efficiency and effectiveness. Efficiency is the time required to find a subset of features and effectiveness is the quality of the subset of features [8]. Depending upon these two factors, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper. The FAST algorithm involves two processes. In the first process, features are divided into clusters using graph-theoretic clustering methods. In the second process, the target classes are selected from each cluster to form a subset of features. Features in different clusters are relatively independent, and thus the clustering-based technique can produce a subset of useful and independent features [9]. The Minimum-Spanning Tree (MST) clustering method is used to evaluate the efficiency of FAST through an empirical study. FAST and many feature selection algorithms have been compared using some experiments. This process is done on four types of well-known classifiers, such as the probability based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The results demonstrate that the FAST not only produces smaller subsets of features but also improves the performances of these classifiers.

The embedded methods involve feature selection as a part of the training process and are usually specific to given learning algorithms. Thus it is more efficient than the other three categories. Traditional machine learning algorithms stand as the best example of embedded approaches. In wrapper methods the generality of the selected features is limited, and the computational complexity is large. In filter methods, even though the computational complexity is very low, the accuracy of the learning algorithms is not guaranteed. The hybrid method makes use of combining the filter and wrapper methods. The main motive of this method is to achieve the best performance with a particular learning algorithm having the similar time complexity of the filter method.

Feature subset selection is the process of identifying and removing all the possible irrelevant and redundant features. This is because an irrelevant feature does not predictive accuracy (i.e., the results cannot be predicted) and redundant features do not redound to getting a better predictor. There are many feature subset selection algorithms, in which some algorithms can effectively eliminate irrelevant features, but it cannot handle redundant features whereas some other techniques such as FAST not only eliminates the irrelevant but also handles redundant features [12]. The feature subset selection research ultimately focuses on searching the relevant features. The best example is Relief. It can weigh each feature based on its ability to discriminate instances. This process is carried out under different targets depending upon the distance-based criteria function. However, Relief cannot remove redundant features as two predictive but the highly correlated features can be highly weighted. Relief-F is an extension of Relief, which works with noisy and incomplete data sets. It deals with multiclass problems, but still, cannot identify redundant features.

III. APRIORI ALGORITHM

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases [3]. This algorithm can be used to find associations between different data sets. It is otherwise known as "Market Basket Analysis." The transaction is referred to as the number of items in each data set. The final result that Apriori algorithm will generate is the sets of rules that tells how often items are contained in sets of data.

The Apriori Algorithm is proposed to find the frequent items in a given data set using the ant monotone constraint. Since Apriori is used for mining frequent item set for Boolean association rules, it is referred to as an influential algorithm in market basket analysis. The general meaning of the Apriori algorithm is that this algorithm uses a prior knowledge of frequent item set properties. Apriori uses an iterative approach which is known as a level-wise search. In this type of search k item set are used to explore (k+1) item set. This algorithm contains some passes over the database. During the first pass k, initially, the algorithm finds the set of frequent item set Lk of length k which satisfies the minimum support requirement. Apriori is designed to operate on databases containing transactions. Each data set contains a number of items and those items are known as transaction. The final result of Apriori algorithm is a set of rules which shows how often items are contained in data set.

IV. FAST ALGORITHMS FOR MINING ASSOCIATION RULES

The main issue of finding association rules between items in a large database depends on sales transactions [1]. Here two algorithms are proposed to overcome this issue. These algorithms totally differ from the other known algorithms such that these algorithms outperform the known algorithms by factors ranging from small and large problems from three to more than an order of magnitude. We also present a hybrid algorithm which combines the best features of the two proposed algorithms called Apriori Hybrid. Several experiments have proved that Apriori Hybrid scales linearly with the total number of transactions. It also has excellent scale-up properties on the transaction size and the number of items in the database.

To extract all the possible association rules, two new algorithms are presented such as Apriori and AprioriTid. These algorithms are compared to the other known algorithms, the AIS and SETM algorithms [10]. The experimental results are presented here which shows that the proposed algorithms always outperform AIS and SETM. The performance gap has been drastically increased on the problem size, and thus the factors range from small and large problems from three to more than an order of magnitude.

The issue of privacy preserving data mining is addressed with the help of the following scenario. Considering two parties who own confidential databases wish to run a data mining algorithm on the union of their databases, without exposing any unwanted information. This work is mainly based on the need to protect privileged information and enabling its use for research or other activities. The above problem is a specific example of secure multi-party computation. This problem can be solved using known generic protocols. The data mining algorithms are typically complex, and the input consists of massive data sets. The known generic protocols cannot be used practically, and therefore more efficient protocols are required. We focus on the problem of decision tree learning with the popular ID3 algorithm. Our protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth.

V. METHODOLOGY

A. Normalized FP-growth

The initial process of this algorithm is to analyze all the possible frequent patterns and then to consider the item set in which it attains the frequency of the threshold. To improve the utility and privacy trade-off, a method is being introduced known as a novel smart splitting method which transforms the possible threats to individual privacy in the pre-processing phase. Differential privacy has been proposed as a way to address such problem which offers strong theoretical guarantees on the released data privacy. It does not make any assumption about an attacker's background knowledge. This differential privacy assures that the output of computation is not sensitive to changes in any individual record and thus it restricts the leakage of privacy through results. Many different algorithms have been introduced for mining frequent item set. The Apriori and FP-growth are the two most prominent ones. Here Apriori algorithm is a candidate set generation and test algorithm, and it uses breadth-first search technique. The appealing features of FP-growth motivate us to design a differentially private FIM algorithm based on the FP-growth algorithm. In this work, we present a practical differentially private FIM algorithm which offers high efficiency rather than achieving high data utility and a high degree of privacy. Since several algorithms of this type have been proposed, many of us are not aware of the existing studies which satisfy all these requirements at a time.

The processes of NFP Growth algorithm are as follows:

1) Initialization.
2) Generation of candidate sets.
3) Local Pruning by reducing the decision tree size.
4) Combining the candidate item set.
5) Calculation of local support and confidence.
6) Publishing the final results.

The resulting demands inevitably bring new challenges. It has been shown that the utility-privacy trade-off can be improved by limiting the length of transactions.

VI. THE ROUND COMPLEXITY OF FREQUENCY ANALYSIS

Just make an assumption that there is a network of three or more players in which each player have some private input. These players are in need to compute specific function of these private inputs in such a way that it protects the privacy of each participant's contribution. Here, it is not necessary for the players to be trusted to do as they are instructed. The resources provided for the players to accomplish their goal are communication i.e., transferring messages to one another in private, or to broadcast messages. The broadcasting can be done in two ways: either to the community as a whole or a local computation. Different protocols have been emerged to solve the problem of multiparty secure function evaluation. Building on Yao's protocol for the case of two players [Ya86], Goldreich, Micali, and Wigderson [GMW87] ordered the first general protocol for this problem, and they provided the paradigm on which a large body of successive work was based. Even though there is an enormous progress, research on secure function evaluation has secured from several serious problems. Since many different protocols have been proposed to overcome this issue; the exact accomplishment has not yet understood. In simple words, it can be said that there is no assurance that these protocols will work effectively in the future. It has been said that these protocols and its underlying techniques have no accepted definitions in this field. Also, the protocols for multiparty secure function evaluation are not efficient. This is because that they require infinite communication rounds. These issues are addressed and a new protocol is designed which improves the complexity

requirement for this task. It entirely divorces the computational complexity of the function. Using this technique, we can overcome the problems with a constant number of rounds of interaction. The result of the new protocol assumes that there is a one-way function, and thus the majority of the participants to the protocol behave correctly.

VII. CONCLUSION

The important issue in generating association rule in the database environment is to analyze frequency in which no site owner is interested to provide database or local frequent item set or support value to other site owners. But generally, all owners wish to access mined result by providing their participation indirectly in the mining process. The problem of knowledge gathering using association rule mining when the database is distributed horizontally among n number of the dataset with a trusted party is considered and analyzed using FP-Growth algorithm.

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.

[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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] R. Fagin, M. Naor, and P. Winkler, "Comparing Information without Leaking It," Comm. ACM, Vol. 39, pp. 77-85, 1996.

[9] M. Freedman, Y. Ishai, B. Pinkas, and O. Feingold, "Keyword Search and Oblivious Pseudorandom Functions," Proc. Second Int'l Conf. Theory of Cryptography (TCC), pp. 303-324, 2005.

[10] 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.

[11] H. Grosskreutz, B. Lemmen, and S. Reuping, "Secure Distributed Subgroup Discovery in Horizontally Partitioned Data," Trans. Data Privacy, Vol. 4, No. 3, pp. 147-165, 2011.

[12] W. Jiang and C. Clifton, "A Secure Distributed Framework for Achieving k-Anonymity," The VLDB J., Vol. 15, pp. 316-333, 2006.

[13] 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.

[14] M. Kantarcioglu, R. Nix, and J. Vida, "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.

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