Horizantal DistributionAssociation Rule Mining

P. Dhana Lakshmi
Assistant Professor
Department of CSSE
Sree Vidyanikethan Engineering College
A. Rangampet

C.Ravindra Murthy
Assistant Professor
Department of EIE
Sree Vidyanikethan Engineering College
A.Rangampet

Abstract: Association rule mining (ARM) has become one of the core data mining tasks. ARM is an undirected or unsupervised data mining technique which works on variable length data, and produces clear and understandable results. Association rule mining is an active data mining research area.ARM algorithms provides to a centralized environment. The aim of Association Rule Mining is to detect relationships or patterns between specific values of categorical variables in large data sets. The main idea of association rule mining in the existing algorithm is to partition the attribute values into Transaction patterns. Basically, this technique enables analysts and researchers to uncover hidden patterns in large data sets. Here the pre-processed data is stored in such a way that online rule generation may be done with a complexity proportional to the size of the output. An optimized algorithm for online rule generation is used in the existing model with adjacency lattice. This algorithm generates all the essential rules and no rule is missing. With the motivation gained from the online rule generation model here we propose a novel Horizontal distributed Association rule mining algorithm for Parallel distributed datasets. ODAM is a distributed algorithm for geographically distributed data sets that reduces communication costs. This algorithm aims to generate rules from different data sets spread over various geographical sites, hence, they require external communications throughout the entire process.

Keywords: Adjacency Latice, Association Rule, Data Mining, Support, Privacy, Item set

Introduction:

Data mining is a research area that investigates the automatic extraction of previously unknown patterns from large amounts of data. It is well documented that this recent advances in data collection without limits explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking.

Privacy preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they find in data privacy. The main consideration in privacy preserving data mining is twofold. First, sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database

by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy, as we will indicate.

The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem. In this report, we provide a classification and an extended description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining.

In all cases when data mining is applied in the context of personal data, basic data and data mining results have to be collected, stored and processed in compliance with data protection legislation. This results in responsibilities for data controllers, technical operators and others involved in those business or governmental processes where data mining plays a role. In this article a brief overview of the state-of-the-art in PPDM and some current suggestions for proceeding towards standardization in PPDM are summarized. To illustrate these considerations, scoring practice in the financial sector is used as an example. Though this example certainly does not demonstrate all aspects possibly relevant in the area of data mining, it has been analyzed from the perspective of recent data protection developments.

Related works:

Privacy Preserving Data Mining is an area where two parties having private databases wish to cooperate by computing a data mining algorithm on the union of their database[3]. Since the databases are confidential, neither party is willing to divulge any of the contents to the other. It shows how the involved data mining problem of decision tree learning can be efficiently computed, with no party learning anything other than the output itself.

Data mining is a recently emerging field, connecting the three worlds of Databases, Artificial Intelligence and Statistics. The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this data is negligible if "meaningful information" or "knowledge" cannot be extracted from it. Data mining, otherwise known as knowledge discovery, attempts to answer this need. In contrast to standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses[1]. As a field, it has introduced new concepts and algorithms such as association rule learning. It has also applied known machine-learning algorithms such as inductive-rule learning (e.g., by decision trees) to the setting where very large databases are involved[14]. Data mining techniques are used in business and research and are becoming more and more popular with time.

Online association rule mining can be applied which helps to remove redundant rules and helps in compact representation of rules for user. Here a new and more optimized algorithm has been proposed for online rule generation. The advantage of this algorithm is that the graph generated in the algorithm has less edge

as compared to the lattice used in the existing algorithm. The use of non redundant association rules help significantly in the reduction of irrelevant noise in the data mining process.

This graph theoretic approach, called adjacency lattice is crucial for online mining of data. The adjacency lattice could be stored either in main memory or secondary memory[15]. The idea of adjacency lattice is to pre store a number of large item sets in special format which reduces disc I/O required in performing the query.

A key problem that arises in any en masse collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual gain. A key utility of large databases today is research, whether it is scientific or economic and market oriented. Thus, for example, the medical field has much to gain by pooling data for research; as can even competing businesses with mutual interests[12]. Despite the potential gain, this is often not possible due to the confidentiality issues which arise.

Data mining results rarely violate privacy, as they generally reveal high-level knowledge rather than disclosing instances of data. However, the concern among privacy advocates is well founded, as bringing data together to support data mining makes misuse easier[10]. The problem is not data mining, but the way data mining is done.

It is clear that any reasonable solution must have the individual parties do the majority of the computation independently. This solution is based on this guiding principle and in fact. The remark that a necessary condition for obtaining such a private protocol is the existence of a (non-private) distributed protocol with low communication complexity[5].

Existing Algorithms:

Online association rule mining can be applied which helps to remove redundant rules and helps in compact representation of rules for user. In this paper, a new and more optimized algorithm has been proposed for online rule generation. The advantage of this algorithm is that the graph generated in our algorithm has less edge as compared to the lattice used in the existing algorithm. The use of non redundant association rules help significantly in the reduction of irrelevant noise in the data mining process.

Online generation of the rules deals with the finding the association rules online by changing the value of the minimum confidence value. Problems with the existing algorithm is that the lattice has to be constructed again for all large item sets, to generate the rules, which is very time consuming for online generation of rule. The number of edges would be more in the generated lattice as we have edges for a frequent item set to all its supersets in the subsequent levels. A weighted directed graph has been constructed and depth first search has been used for rule generation. In the proposed algorithm, online rules can be generated by generating adjacency matrix for some confidence value and the generating rules for confidence measure higher than that used for generating the adjacency matrix.

This graph theoretic approach, called adjacency lattice is crucial for online mining of data. The adjacency lattice could be stored either in main memory or secondary memory. The idea of adjacency lattice is to pre store a number of large item sets in special format which reduces disc I/O required in performing the query.

An itemset *X* is said to be adjacent to an itemset *Y* if one of them can be obtained from the other by adding a single item. Specifically, an itemset *X* is said to be a parent of the itemset *Y*, if *Y* can be obtained from *X* by adding a single item to the set *X*. It is clear that an itemset may possibly have more than one parent and more than one child. In fact, the number of parents of an itemset *X* is exactly equal to the cardinality of the set *X*.

The three steps involved are:

1. Generation of adjacency lattice
2. Online Generation of Item sets
3. Rule Generation

1. Generation of adjacency lattice:

The Adjacency lattice is created using the frequent item sets generated using any standard algorithm by defining some minimum support. This support value is called primary threshold value. The item sets obtained above are referred as pre-stored item sets**,** and can be stored in main memory or secondary memory. This is beneficial in the sense that we need not to refer dataset again and again from different value of the min. support and confidence given by the user.

The adjacency lattice L is a directed acyclic graph. An itemset X is said to be adjacent to an itemset Y if one of them can be obtained from the other by adding a single item. The adjacency lattice L is constructed as follows: Construct a graph with a vertex v(I) for each primary itemset I. Each vertex I has a label corresponding to the value of its support. This label is denoted by S(I). For any pair of vertices corresponding to itemsets X and Y, a directed edge exists from v(X) to v(Y) if and only if X is a parent of Y. Note that it is not possible to perform online mining of association rules at levels less than the primary threshold.

2. Online Generation of Item sets:

Once we have stored adjacency lattice in RAM. Now user can get some specific large item sets as he desired. Suppose user want to find all large item sets which contain a set of items I and satisfy a level of

minimum support s, then there is need to solve the following search in the adjacency lattice. For a given itemset I, find all item sets J such that v(J) is reachable from v(I) by a directed path in the lattice L, and satisfies $S(J) \geq s$.

3. Rule Generation:

Rules are generated by using these pre-stored item sets for some user defined minimum support and minimum confidence. This dataset has five transactions and five item sets.

Proposed Algorithm:

Data mining is a technology to explore, analyze and finally discovering patterns from large data repository. Association rule mining is a technique in data mining which is used to describe relations among items in transactions. There are two types of database environments exist namely centralized and distributed.

- Centralized
- Distributed

Centralized: Centralized database is a database in which data is stored and maintained in a single location. This is the traditional approach for storing data in large enterprises.

Distributed: A distributed data base is a data base in which portions of the data base are stored on multiple locations within a network.

Association rule mining (ARM) is an active data mining research area. However, most ARM algorithms provide supply to a centralized environment. In contrast to previous ARM algorithms, ODAM is a distributed algorithm for geographically distributed data sets that reduce communication costs.

Modern organizations are geographically distributed. Typically, each site locally stores its ever increasing amount of day-to-day data. Using centralized data mining to discover useful patterns in such organizations data is not always feasible because merging data sets from different sites into a centralized site causes huge network communication costs. Data from these organizations are not distributed over various locations but also vertically fragmented, making it difficult if not impossible to combine them in a central location.

Algorithm:

Step1: Load the data sets.
Step2: Enter number of partitions.
Step3: Perturbation takes place i.e shuffling of data takes place.
Step4: Partitions the data horizontally and is distributed to servers.
Step5: Adds noise to transaction count .
Step6: Distributes data to mining nodes.
Step7: Performs frequent item set mining on given partitions using adjacency lattice approach.
Step8: Find TLS of each frequent items.

Step9: Upon receiving all frequent item sets, finds all frequent item sets with given support.
Step10: Display the global frequent item sets and actual support.
Step11: Generation of rules by using global frequent item sets and their actual support and confidence.
Step12: Display all generated rules.

Results:

PERFORMANCE EVALUATION

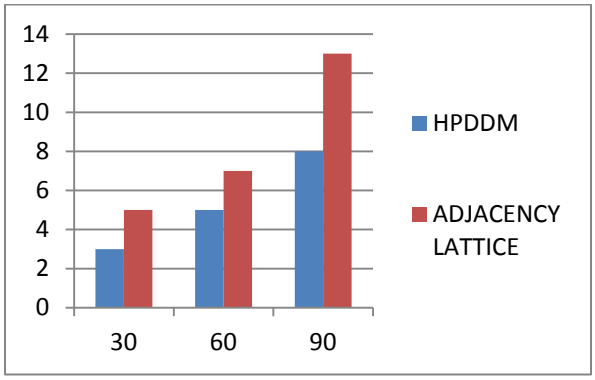

FIGURE.1: PERFORMANCE EVALUATION OF HORIZONTAL DISTRIBUTED ASSOCIATION RULE MINING

x-axis represents the number of transactions.
y-axis represents the time taken in seconds

To calculate the weight of the edge between itemset Xand itemset Y, where (X-Y) = 1-itemset, calculate the value support(X)/support(Y) if this value is >= minimum confidence then we can have an edge between the itemset X

and the itemset Y and this edge will have weight =support(X)/support(Y).

From figure 5.1 it can be observed that the time taken for the generation of rules is very less through horizontal distributed association rule mining for parallel data sets as compared to adjacency lattice. Hence we can say that it is efficient in terms of time when compared to centralized based approach adjacency lattice.

CONCLUSION AND FUTURE WORK

A novel connected perturbation approach to perturbate the attribute position that protects the information about the role of that attribute in dataset and also achieves perturbation in resultant item sets. Most of the solutions currently available in recent literature either not compatible to distributed data mining or fail to avoid data leaks under nexus of one or more data mining node authorities. In this regard the model that proposed here is perturbating actual dataset in connected manner. The proposed model that referred as “Hierarchical Privacy Preserving Distributed Frequent Item set Mining Over Horizontally Distributed-Dataset (HPPDFIM-HD)” is perturbating the actual item frequency, position and field id in connected manner, also perturbating the actual horizontal partition size by adding records as Gaussian noise. In future work, it is recommended to develop perturbation technique that perturbate even attribute count.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, Chile: VLDB, Sept. 12-15 1994, pp. 487-499.

[2] D. W.-L. Cheung, J. Han, V. Ng, A. W.-C. Fu, and Y. Fu, "A fast distributed algorithm for mining association rules," in Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (PDIS'96). Miami Beach, Florida, USA: IEEE, Dec. 1996, pp. 31-42.

[3] D. W.-L. Cheung, V. Ng, A. W.-C. Fu, and Y. Fu, "Efficient mining of association rules in distributed databases," IEEE Trans. Knowledge Data Eng., vol. 8, no. 6, pp. 911-922, Dec. 1996.

[4] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in Proceedings of the 2000 ACM SIGMOD Conference on Management of Data. Dallas, TX: ACM, May 14-19 2000, pp. 439-450.

[5] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Santa Barbara, California, USA: ACM, May 21-23 2001, pp. 247-255.

[6] A. Ev?mievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26 2002, pp. 217-228.

[7] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in Proceedings of 28th International Conference on Very Large Data Bases. Hong Kong: VLDB, Aug. 20-23 2002, pp. 682-693.

[8] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Advances in Cryptology - CRYPTO 2000. Springer-Verlag, Aug. 20-24 2000, pp. 36-54.

[9] O. Goldreich, "Secure multi-party computation," Sept. 1998, (working draft).

[10] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26 2002, pp. 639-644.

[11] A. C. Yao, "How to generate and exchange secrets," in Proceedings of the 27th IEEE Symposium on Foundations of Computer Science. IEEE, 1986, pp. 162-167.

[12] I. Ioannidis and A. Grama, "An efficient protocol for yao's millionaires'
problem," in Hawaii International Conference on System Sciences (HICSS 36), Waikoloa Village, Hawaii, Jan. 6-9 2003.

[13] O. Goldreich, "Encryption schemes," Mar. 2003, (working draft).

[14] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," Communications of the ACM, vol. 21, no. 2, pp. 120-126, 1978.

[15]C. Farkas and S. Jajodia. The Inference Problem: A Survey. In SIGKDD Explorations, 4(2). 6-11, December 2002.

[16] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin and M. Y. Zhu. Tools for Privacy Preserving Distributed Data Mining. In SIGKDD Explorations, 4(2). 28-34 December 2002.

[17] I. Dinur, K. Nissim. Revealing information while preserving privacy. Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. California. 2003. pp.202-210.