Secure Mining of Association Rules in Databases using Distributed Database Algorithm

A. Ranjith Kumar
M.Phil. Research Scholar
Department of Computer Science, Morappur Kongu College of Arts &Science

OPEN ACCESS

Volume : 6

Special Issue : 1

Month : September

Year: 2018

ISSN: 2321-788X

Impact Factor: 3.025

Citation:
Ranjith Kumar, A. (2018). Secure Mining of Association Rules in Databases using Distributed Database Algorithm. *Shanlax International Journal of Arts, Science and Humanities*, 6(S1), pp.61–66.

DOI:
https://doi.org/10.5281/zenodo.1410981

Abstract

We propose a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Distribute Database Algorithm. Our protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm. Which is an unsecured distributed version of the Apriority algorithm? The main ingredients in our protocol are two novel secure multi-party algorithms — one that computes the union of a private subsets that each of the interacting players holds, and another that tests the inclusion of an element held by one player in a separation held by another. Our protocol offers enhanced privacy concerning the protocol. Also it is simpler and is significantly more efficient regarding communication rounds, communication cost, and computational cost.

Introduction

Overview

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one some analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases.

This includes are following:

1. Operational or transactional data such as sales, cost, inventory, payroll, and accounting.

2. Nonoperational data, such as industry sales, forecast data, and macro-economic data
3. Meta data - data about the data itself, such as logical database design or data dictionary definitions

Information Data Mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

Generally, any of four the types of relationships are sought:

1. Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
2. Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
3. Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
4. Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data Mining consists of Five Major Elements

1. Extract, transform and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

Different Levels of analysis are available

1. Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2. Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
3. Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi-square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
4. Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.
5. Rule induction: The extraction of useful if-then rules from data based on statistical significance.
6. Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Literature Review

Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data

Data mining can extract important knowledge from large data collections – but sometimes these collections are split among various parties. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. This paper addresses secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared while adding little overhead to the mining task.

Privacy-Preserving Mining of Association Rules

We present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve the privacy of individual transactions. While it is feasible to recover association rules and preserve privacy using a straightforward \uniform" randomization, the discovered rules can, unfortunately, be exploited to and privacy breaches. We analyze the nature of privacy breaches and propose a class of randomization operators that are much more effective than uniform randomization in limiting the breaches. We derive formulae for an unbiased support estimator and its variance, which allow us to recover item set supports from randomized datasets, and show how to incorporate these formulae into mining algorithms. Finally, we present experimental results that validate the algorithm by applying it on real datasets.

An Efficient Approximate Protocol for Privacy Preserving Association Rule Mining

The secure scalar product (or dot product) is one of the most used sub-protocols in privacy-preserving data mining. Indeed, the dot product is probably the most common sub-protocol used. As such, a lot of attention has been focused on coming up with secure protocols for computing it. However, an inherent problem with these protocols is the extremely high computation cost especially when the dot product needs to be carried out over large vectors. This is quite-common in vertically partitioned data and is a real problem. In this paper, we present ways to efficiently compute the approximate dot product. We implement the dot-product protocol and demonstrate the quality of the approximation. Our dot product protocol can be used to securely and efficiently compute association rules from data vertically partitioned between two parties.

Keyword Search and Oblivious Pseudorandom Functions

We study the problem of privacy-preserving access to a database. Particularly, we consider the problem of privacy-preserving keyword search (KS), where records in the database are accessed according to their associated keywords and where we care for the privacy of both the client and the server. We provide efficient solutions for various settings of KS, based either on specific assumptions or general primitives (mainly oblivious transfer). Our general solutions rely on a new connection between KS and the oblivious evaluation of pseudorandom functions (OPRFs). We, therefore, study both the definition and construction of OPRFs and, as a corollary, give improved constructions of OPRFs that may be of independent interest.

Privacy-Preserving Mining of Association Rules

We present a framework for mining association rules from transactions consisting of the categorical items where the data has been randomized to preserve the privacy of individual transactions. While it is feasible to recover association rules and preserve privacy using a straightforward ‘‘uniform’’ randomization, the discovered rules can, unfortunately, be exploited to find privacy breaches. We analyze the nature of privacy breaches and propose a class of randomization operators that are

much more effective than uniform-randomization in limiting the breaches. We derive formulae for an unbiased support estimator, and its variance, which allow us to recover item set supports from randomized datasets, and show how to incorporate these formulae into mining algorithms. Finally, we present experimental results that validate the algorithm by applying it on real datasets.

Secure Mining of Association Rules In Horizontally Distributed

We propose a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton. Our protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm of Cheung et al., which is an unsecured distributed version of the Apriority algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms one that computes the union of a private subsets that each of the interacting players holds, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy concerning the protocol in [18]. Also, it is simpler and is significantly more efficient regarding communication rounds, communication cost, and computational cost.

Privacy-Preserving Data Mining

In this paper, we address the issue of privacy preserving data mining. Specifically, we consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes. The above problem is a specific example of secure multi-party computation and as such, can be solved using known generic protocols. However, data mining algorithms are typically complex and, furthermore, the input usually consists of massive data sets. The generic protocols in such a case are of no practical use, and therefore more efficient protocols are required. We focus on the problem of decision tree learning with the popular ID3 algorithm.

Project Description

Module Description

Modules

1. Privacy Preserving Data Mining
2. Distributed Computation
3. Frequent Item sets
4. Association Rules

Modules Description

Privacy Preserving Data Mining

One, in which the data owner and the data miner are two different entities, and another, in which the data is distributed among several parties who aim to jointly perform data mining on the unified corpus of data that they hold. In the first setting, the goal is to protect the data records from the data miner. Hence, the data owner aims at anonym zing the data before its release. The main approach in this context is to apply data perturbation. The idea is that. Computation and communication costs versus the number of transactions N the perturbed data can be used to infer general trends in the data, without revealing original record information. In the second setting, the goal is to perform data mining while protecting the data records of each of the data owners from the other data owners. This is a problem of secure multiparty computation. The usual approach here is cryptographic rather than probabilistic.

Distributed Computation

We compared the performance of two secure implementations of the FDM algorithm Section In the first implementation (denoted FDM-KC), we executed the unification step using Protocol DDA, where the commutative cipher was 1024-bit RSA in the second implementation (denoted FDM) we used our Protocol DDA, where the keyed-hash function was HMAC.

In both implementations, we implemented the DDA algorithm in a secure manner. We tested the two implementations concerning three measures:

1. The total computation time of the complete protocols (DDA) over all players. That measure includes the Apriority computation time, and the time to identify the globally s-frequent item sets, as described in later.
2. The total computation time of the unification protocols only (DDA) over all players.
3. The total message size. We ran three experiment sets, where each set tested the dependence of the above measures on a different parameter: • N — the number of transactions in the unified database.

Frequent Item Sets

We describe here the solution that was proposed by DDA. They considered two possible settings. If the required output includes all globally s-frequent item sets, as well as the sizes of their supports, then the values of $\Delta(x)$ can be revealed for all. In such a case, those values may be computed using a secure summation protocol, where the private addend of Pm is $ppm(x) - sNm$. The more interesting setting, however, is the one where the support sizes are not part of the required output. We proceed to discuss it.

Association Rules

Once the set Fs of all s-frequent item sets is found, we may proceed to look for all (s, c)-association rules (rules with support at least sN and confidence at least c). To derive from Fs all (s, c)-association rules in an efficient manner we rely upon the straightforward lemma.

Conclusion

In this project devise a protocol for secure mining of association rules in horizontally partitioned distributed databases. The protocol is more efficient the current leading K and C protocol. The main ingredients of this protocol are two novel secure multiparty algorithms in which these two main operations are union and intersection. The protocol exploits the fact that the underlying problem is of interest only if the number of players are more than two. The direction to future work is to devise an efficient protocol for inequality verifications that uses the existence of semi-honest third party and another in Implementation of the techniques to the problem of distributed association rule mining in a vertical setting.

References

Agrawal, R., & Srikant, R, "Privacy-preserving data mining," SIGMOD Conference, pages 439–450, 2000. 33.

Evfimievski, AV, Srikant, R, Agrawal, R., & Gehrke, J, (2002). "Privacy preserving mining of association rules," In KDD, pages 217–228.

Evfimievski, AV, Srikant, R, Agrawal, R., & Gehrke, J. (2002). "Privacy preserving mining of association rules," In KDD, pages 217–228.

Freedman, M, Ishai, Y, Pinkas, B., & Reingold, O, (2005). "Keyword search and oblivious pseudorandom functions", In TCC, pages 303–324.

Kantarcioglu, M, Nix, R., & Vaidya, J, (2009). "An efficient approximate protocol for privacy preserving association rule mining," In PAKDD, pages 515– 524.

Kantarcioglu, M ., & C. Clifton, C (2004). "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge and Data Engineering, 16:1026–1037.

Tamir TASSA (2013). "Secure Mining of Association Rules in Horizontally Distributed Databases," IEEE transactions on knowledge and data engineering, 2013.