International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Privacy Based Association Rules in Secure Horizontal Database

Zameena R[1], Nitha L Rozario[2]

[1]M. Tech Student, Marian Engineering College, Kerala University, Trivandrum, Kerala, India

[2]Assistant Professor, Dept. of CSE, Marian Engineering College, Trivandrum, Kerala, India

Abstract: ***Privacy in data mining prevents the parties from directly sharing the data, and some types of information about the data. We introduce a protocol for secure mining of association rules in horizontally distributed databases using FDM, which an unsecured distributed version of the Apriori algorithm. We extended our work with Horizontal Aggregation that give rise to multiple row output, which transform rows to column using CASE, SPJ or PIVOT operators depending on the input. In order to prepare real world datasets that are very much suitable for data mining operations, we explored horizontal aggregations by developing constructs in the form of operators such as CASE, SPJ and PIVOT. Instead of single value, the horizontal aggregations return a set of values in the form of a row. The proposed system is user friendly as users are never expected to write queries we introduced security and privacy through cryptography and level based slicing. Level Based Slicing introduces different techniques such as Generalized Data, Bucketized Data, Multiset-based Generalization Data, One-attribute-per-Column Slicing Data, Sliced Data dimensionality of the data and preserves better utility than generalization and bucketization. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying. Our result shows that proposed method have higher performance other sequential algorithms. Customers can dynamically send request to the server either to apply privacy and security. In privacy, server can apply five types of slicing methods. In security server can apply FDM, Encryption and decryption, Hash function generate key for encryption and decryption. Experimental Analysis can be comparing with the time taken to evaluate the privacy preserving methods and Enhanced FDM Algorithm.***

Keywords: privacy preserving data mining, Advanced Encryption Standard, Association Rules, Level Based Slicing, Horizontal Aggregation

1. Introduction

Distributed computing plays an important role in the Data Mining process. Data Mining often requires huge amounts of resources in storage space and computation time. We extended our work with Horizontal Aggregation that give rise to multiple row output, which transform rows to column using CASE, SPJ or PIVOT operators depending on the input. We explored horizontal aggregations by developing constructs in the form of operators such as CASE, SPJ and PIVOT. Instead of single value, the horizontal aggregations return a set of values in the form of a row. The result resembles a multidimensional vector. We have implemented SPJ using standard relational query operations. The CASE construct is developed extending SQL CASE [1]. The PIVOT makes use of built in operator provided by RDBMS for pivoting data. To evaluate these operators, we have developed a web based prototype application and results reveal that the proposed horizontal aggregations are capable of preparing data sets for real world data mining operations. As the proposed constructs are based on SQL, it reduces lot of coding as it is a powerful data retrieval language. The proposed system is user friendly as users are never expected to write queries. Instead, they just make queries in a user-friendly fashion. In modern database where data is stored because of day to day operations, users do not get a chance to use data directly for mining operations. Instead, it has to be transformed to make sense for data mining operations. Generally data from business database is converted, and loaded into some other data sets known as data warehouse. The proposed horizontal aggregations can be used to generate data sets for the purpose of data mining analysis. We introduced security and privacy through cryptography and level based slicing. Level Based Slicing introduce different techniques such as Generalized Data, Bucketized Data, Multiset-based Generalization Data, One-attribute-per-Column Slicing Data, Sliced Data[2].Slicing partitions, the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes... Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. This paper studies the problem of association rules mining in horizontally distributed databases. In the distributed databases, there are several players that hold homogeneous databases which share the same schema but hold information on different entities. The information that we would like to protect in this context is not only individual transactions in the different databases, but also more global information such as what association rules are supported locally in each of those databases [3]

2. Literature Survey

Data aggregation is a process in which information is Gathered and expressed in a summary form, and which is used for purposes such as statistical analysis [1]. Horizontal Aggregations helps building answer sets in tabular form, which in standard form needed by most data mining algorithms. In horizontal aggregation, a new class of aggregations has similar behavior to SQL standard

aggregations, but which produce tables with a horizontal layout. In contrast, standard SQL aggregations that is vertical aggregations which produce tables with a vertical layout. Horizontal aggregations just require a small syntax extension to aggregate functions called in a SELECT statement.

Here, a new class of aggregations introduces that have similar behavior to SQL standard aggregations, but which produce tables with a horizontal layout. In contrast, we call standard SQL aggregations vertical aggregations since they produce tables with a vertical layout [1]. Horizontal Aggregations just require a small syntax extension to aggregate functions called in a SELECT statement.

The paper [3] studied the problem of secure mining of association rules in horizontally partitioned databases. Tamir Tassa proposed here a protocol Fast Distributed Mining algorithm (FDM) for mining of association rules in horizontally distributed databases. The main idea is that the player's finds their locally s-frequent itemsets then the players check each of them to find out globally s-frequent item set. Paper assumes that the players are semi honest; they try to extract information. Hence the player compute the encryption of their private database together by applying commutative encryption .Paper shows that their protocol offers better privacy and is significantly more efficient in terms of communication cost and computational cost while the solution is still not perfectly secure cause it leaks excess information [3].

The paper [4] Proposed data anonymization technique called slicing to improve the current state of threat. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns

Advantages:

- The proposed system ensures the anonymity requirement.
- Improves the accuracy of the system and performance.

2.1 FDM Algorithm

The protocol of [5], as well as ours, is based on the Fast Distributed Mining (FDM) algorithm. It is an unsecured distributed version of the Apriori algorithm. Its main idea is that any *s*-frequent itemset must be also locally *s*-frequent in at least one of the sites. Hence, in order to find all globally *s*-frequent itemsets, each player reveals his locally *s*-frequent itemsets and then the players. Check each of them to see if they are *s*-frequent also globally. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis"
The FDM algorithm proceeds as follows:
1) Initialization
2) Candidate Sets Generation
3) Local Pruning
4) Unifying the candidate itemsets
5) Computing local supports
6) Broadcast Mining Results

3. Method

Data aggregation is a process in which information is gathered and expressed in a summary form, and which is used for purposes such as statistical analysis. Horizontal aggregations helps building answer sets in tabular form, which in standard form needed by most data mining algorithms. In horizontal aggregation, a new class of aggregations it produce tables with a horizontal layout. Horizontal aggregations just require a small syntax extension to aggregate functions called in a SELECT statement.
Horizontal aggregation is evaluated using three fundamental methods:

3.1 CASE, SPJ (Select Project Join) and PIVOT

1) CASE Method
Two basic strategies to compute horizontal aggregations: The first strategy is to compute directly from input table. The second approach is to compute vertical aggregation and save the results into temporary table. Then that table is further used to compute horizontal aggregations

2) SPJ Method
This method is based on the relational operators only. In this method one table is created with vertical aggregation for each column. Then all such tables are joined in order to generate a table containing horizontal aggregations. This method performs Select, Project and Join operation on the fact table

3).PIVOT Method:
The pivot operator is a built-in operator which transforms row to columns. It internally needs to determine how many columns are needed to store the transposed table and it can be combined with the GROUP BY clause

3.2 Privacy Preservation of Data Using Slicing Method

We conduct extensive workload experiments. Our results confirm that slicing preserves much better data utility than generalization. In workloads involving the sensitive attribute, slicing is also more effective than bucketization. In some classification experiments, slicing shows better performance than using the original data. Different Methods are

1) Generalized Data
Generalized Data, in order to perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. This significantly reduces the data utility of the generalized data.

2) Bucketized Data
We show the effectiveness of slicing in membership disclosure protection. For this purpose, we count the number

of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake tuples. Our experiment results show that bucketization does not prevent membership disclosure as almost every tuple is uniquely identifiable in the bucketized data

3) Multiset-based Generalization Data

We observe that this multiset-based generalization is equivalent to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but break the association between them within one bucket.

4) One-attribute-per-Column Slicing Data

We observe that while one-attribute-per-column slicing preserves attribute distributional information, it does not preserve attribute correlation, because each attribute is in its own column. In slicing, one group correlated attributes together in one column and preserves their correlation. For example, in the sliced table shown in Table correlations between Age and Sex and correlations between Zip code and Disease are preserved. In fact, the sliced table encodes the same amount of information as the original data with regard to correlations between attributes in the same column.

5) Sliced Data

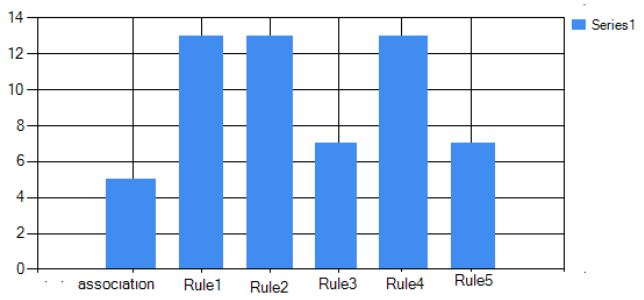
Another important advantage of slicing is its ability to handle high-dimensional data. By partitioning attributes into columns, slicing reduces the dimensionality of the data. Each column of the table can be viewed as a sub-table with a lower dimensionality. Slicing is also different from the approach of publishing multiple independent sub-tables in that these sub-tables are linked by the buckets in slicing. You should use Times Roman of size 10 for all fonts in the paper. Format the page as two columns:

4. Result

The performance of the level based privacy preserved discrimination free data transmission software can be analysed on the basis of the security it provides. This software will provide high security without any information loss issue and can be used in any organization. Analysis can be comparing with the time taken to evaluate the privacy preserving methods and Enhanced FDM Algorithm

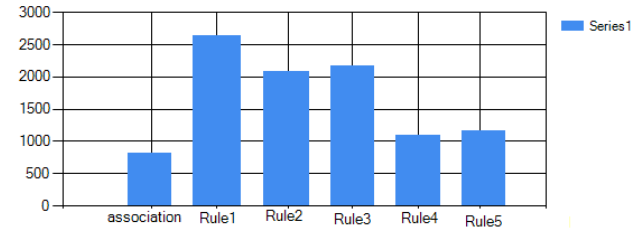
4.1 Rule Analysis

Compare the five Privacy Rules with Association Rule. Rule Analysis compares how many times each rule to use in different customers... Different Rules are Generalized Data, Bucketized Data, Multiset-based Generalization Data, One-attribute-per-Column Slicing Data, Sliced Data and Fast Distributed Mining Algorithm.



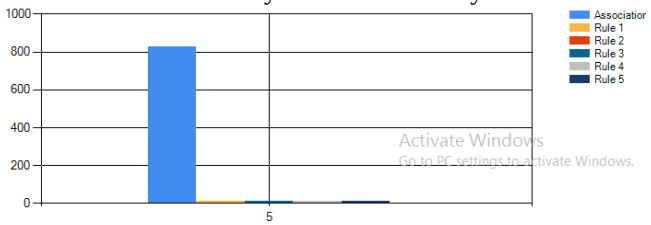
4.2 Time Analysis

Compare the five Privacy Rules with Association Rule. Rule Analysis compares the time to taken for evaluation of each rule. Time can calculate by milliseconds. Different Rules are Generalized Data, Bucketized Data, Multiset-based Generalization Data, One-attribute-per-Column Slicing Data, Sliced Data and Fast Distributed Mining Algorithm



4.3 Overall Analysis

Here we calculate the overall analysis by taken the combination of rule Analysis and Time Analysis



5. Conclusion

The proposed software is dynamic in nature and thus can be installed in any organization and can handle any databases. According to the degree of privacy required to each agent, the different privacy preservation rule sets are applied to it. Data transformation or Data preparation is an important stage of knowledge discovery process. Data transformation yields output which is considered as a better input for data mining process. SQL provides aggregation functions which generate output as a single row and no increase in number of columns. A new class of aggregate function, called horizontal aggregation over multiple databases is used to prepare data sets for data mining analysis from multiple databases. The three methods demonstrated in proposed framework SPJ, CASE and PIVOT, overcome the shortcomings of SQL vertical aggregation. SPJ and CASE methods are traditionally not in-built function or feature of SQL to generate horizontally aggregated columns whereas PIVOT uses an in-built feature of SQL in grouping along aggregation function. The three methods from the framework SPJ, CASE and PIVOT methods allow drop down option for available

column names to generate horizontally aggregated dataset. PIVOT and CASE can be thought of as the fastest methods out of the three. The implementation of SPJ, CASE and PIVOT in this framework reduces complexity of writing SQL code for preparing datasets from multiple databases. The main objective that the software focuses is the correlation between the security and privacy preservation. So, we can conclude that by preserving privacy itself, the increase the security

References

[1] Carlos Ordonez, Zhibo Chen. "Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis," IEEE Trans on Knowledge and Data Engineering,10.1109/TKDE.2011.16, April 2012.
[2] C. Dwork, "Differential Privacy: A Survey of Results," Proc. Fifth Int'l Conf. Theory and Applications of Models of Computation (TAMC), pp. 1-19, 2008
[3] Tamir Tassa,"Secure mining of association rule in horizontally distributed databases", IEEETrans. Knowledge and Data Engg, Vol. 26, no.2, April 2014
[4] Tiancheng Li, Ninghui Li, Senior Member, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" Proc.", IEEETrans. Knowledge and Data Engg, VOL. 24, NO. 3, MARCH 2012
[5] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, Sept. 2004.

Author Profile

Zameena. R, Doing M. Tech degree in CSE Dept , from Marian Engineering College, Kazhakuttom,Trivandrum, Kerala, India in 2014-2016.

Ms. Nitha L Rozario, Assistant Professor, Dept. of CSE, Marian Engineering College, Trivandrum, Kerala, India, She is done M.tech from Tkmit,Ezhukone,Kollam under the cusat university.