Privacy Preserving Combinatorial Function for Multi-Partitioned Data Sets

V.S. Prakash
Assistant Professor
Department of CA

A. Shanmugam
Principal

P. Murugesan
Assistant Professor (SG)
Department of CA

BIT, Sathyamangalam,Tamil Nadu, India.

ABSTRACT

To preserve the electronic fitness records of data mining applications, Data perturbation is used. It is a form of privacy-preserving data mining. An improved amount and range of information stored in databases has direct to an enhancement in the desire for ranked and “best match” queries. Such queries are mainly applicable when dealing with privacy-sensitive information. To facilitate privacy preservation in data mining or machine learning algorithms over horizontally partitioned or vertically partitioned data, many protocols have been proposed using SMC and various secure building blocks.
In this work, we plan to present an effective and efficient cluster based privacy preserving data perturbation technique to mine Multi-partitioned data sets that comprises of both vertical and horizontal data sets which is current demand of e-business and e-commerce data mining environment. To evaluate a trade of between data privacy and transparency of individual’s data, data perturbation technique is presented with validation and authentication. In multi-partitioned data distribution, data perturbation raised ambiguity between vertical and horizontal partitions of the data. To overcome the ambiguity, we plan to introduce divisive k-neighbor clusters for multi-partitioned data sets. The performance is evaluated with bench data sets obtained from popular e-business / e-commerce sites. (Amazon, e-bay etc.,).

Keywords

Multi-partitioned data sets, Privacy preservation, Data perturbation technique.

1. INTRODUCTION

Data perturbation has been termed as a more efficient application of data security in health care than de-identification/re-identification owing to the advanced probability that strikes could take place which link public data sets to original identifiers or subjects. So data perturbation is termed as a firm application for electronic health care data security. The probability distribution approach receives the data and returns it from the similar distribution model or from the distribution itself.

The value distortion approach agitates data by multiplicative or additive noise, or other randomized processes. It is measured to be more efficient than the previous type of perturbation. This approach assembles decision tree classifiers where each element is dispersed random noise from the Gaussian distribution, for instance. By data mining, the unique data distribution is restored from its agitated version. Though, opponents point to the fact that random additive noise can be filtered which can result in privacy compromises. Data perturbation is termed a fairly simple and efficient technique in for protecting sensitive electronic data from unauthorized use.

Without any direct access to the original data sets, Privacy-preserving data mining (PPDM) considers the problems of completing data mining tasks, because the providers declare privacy on their data. The cryptographic techniques from SMC offer solutions which always stipulate high computation cost, particularly on processing volumes of data in data mining applications. Other approaches are supported on data perturbation techniques, which are efficient than the techniques of SMC.

A current progression under computing, communication, and digital storage technologies have allowed unbelievable volumes of data to be available slightly across geographical and administrative boundaries. There is a growing stipulation on mutual mining over the distributed data stores to find the patterns or rules that benefit all of the participants. For instance, various trader stores in the same business section want to group their data together to decide the characteristics of customer purchases. Cancer research institutes in various geographical areas want to collaboratively discover the ecological factors related to certain type of cancer. But, these distributed data sets could also include perceptive information, such as business sales data and patient clinical records. As a result, an imperative challenge for distributed collaborative mining is how to guard each participant’s sensitive information, while still finding useful data models.

The key challenge for applying Combinatorial Function for multi-partitioned data technique is to securely unify the perturbations used by different data providers, while each party still receives adequate privacy assured and the utility of the pooled data is well preserved. The data partitioning process is taken place by different techniques.

2. LITERATURE REVIEW

Partitioning is a process in which the data has been partitioned among different users from the database. The partitioning is made up of two types, horizontally partition, vertically partition. The data sets obtained from these two partitions may vary. In [10] presented a technique for data partitioning using multi classifier systems. The filter based data partitioning is made here. But the scalability of the scheme is very low.

For these two types of partition, [6] presented a technique for vertically partitioned dataset in a privacy preserving manner using SVM Classification. But the classification is done efficiently with the datasets. [1] presented the evaluation for kth element score in vertically partitioned data. The main concern over this technique is that it supports only single partitioned data. [2] presented a gradient descent methods for privacy preservation. It is necessary to share the data among the users in a secure manner. This too supports only for single-partitioned data sets.

The privacy preservation is done with several techniques which have been presented earlier. In [4] proposed a privacy preservation process using nearest neighbor search. In [8] presented a method for privacy preservation i.e., to protect the shared data from third parties. The Linear Fisher Discriminant Analysis for privacy preservation is used to protect the data but the reliability of the system fails.

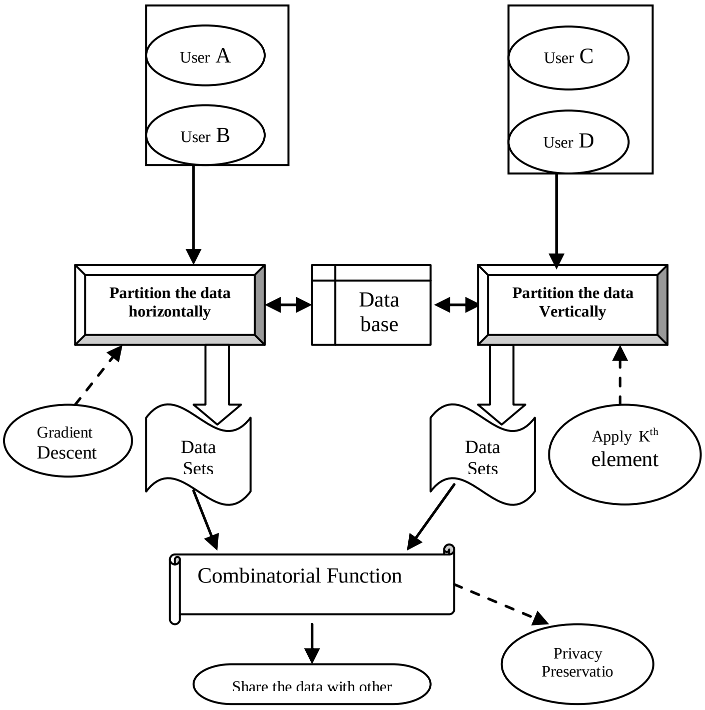
In [9] presented a technique for privacy preservation using Genetic Algorithms for Rule Discovery. The rules have been generated with Genetic Algorithms for the protection of shared datasets. In [12] proposed privacy preservation for the same sets but with distributed sets of data. The authenticity of the data may be lost here due to the distributed approach.

To overcome all the issues arised with single partitioned data, here we are going to propose a technique termed privacy preserving combinatorial function for multi-partitioned datasets. The scalability and reliability of the datasets will be high while applying the technique for multi-partitioned data.

3. COMBINATORIAL FUNCTION FOR MULTI-PARTITIONED DATA SETS

The critical design of data perturbation is to modify the data so that valid individual data values cannot be improved, while preserving the value of the data for statistical summaries. Since the data doesn't react the real values of private data, even if a data item is linked to an individual that individual's privacy is not violated.

Multi-partitioned data set consists of data which have been divided from any logical database. The data has been partitioned into two types: horizontally data partition and vertically data partition. Horizontal partitioning engages setting diverse rows into diverse tables. For instance, customers with codes less than 50000 are amassed in CustomersEast, whereas customers with codes greater than or equal to 50000 are amassed in CustomersWest. The partition tables are then CustomersEast and CustomersWest, which have been combined to present a absolute view of all customers.



3.1 Architecture Diagram for combinatorial function in multi-partitioned data

A general outline of vertical partitioning is to divide dynamic data from static data in a table where the dynamic data is not used as often as the static. Generating a view for the two tables re-establish the unique table with a performance penalty, though the performance will increase when accessing the static data e.g. for statistical analysis. It is necessary to develop a privacy preserving scheme for both horizontal and vertical partitioned data sets using data perturbation technique. As shown in figure 3.1, User A and User B wants to partition the data in the database as horizontally partitioned data sets. After partition the data from the database, they have some sets of data to share it with other users in a safe and secure manner. That is, no other third party should involve in this sharing data between the users. Then User C and User D partition the data from the database in a vertical manner. They received some sets of vertically partitioned data sets.

To share the data sets with different users along with privacy preservation, the proposed work presented a technique named Data Perturbation technique which is used to share the data from different users and unite all those data to get one complete true data sets. The combinatorial function is used to preserve the data sets which are to be shared among the users and it does not allow the third party members to seize the data. The proposed combinatorial function will allow the users to share both the horizontal and vertical partitioned data sets to share with different users.

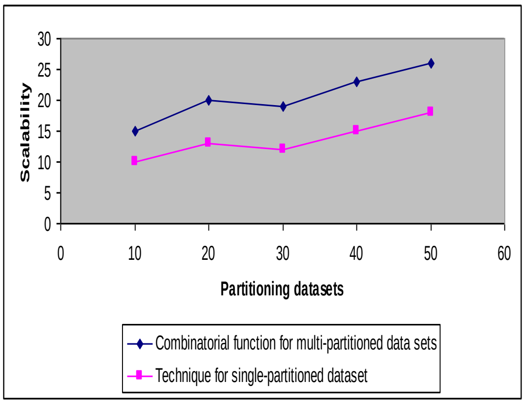
4. PERFORMANCE EVALUATION

The proposed combinatorial function for multi-partitioned data has been implemented in Java. The experiments were run on an Intel P-IV machine with 2 GB memory and 3 GHz dual processor CPU. We are going to compare the privacy preserving partitioning data using combinatorial function for multi-partitioned data with an existing technique for individual partition data sets.

To the best of our knowledge, there is no technique that can perform the privacy preservation for multi-partitioned data sets using combinatorial function. While using combinatorial function for multi-partitioned datasets, the scalability of the products/services is improved. The validity of the data set to be shared remains same from the beginning of the partitioning process using combinatorial function.

5. RESULTS AND DISCUSSION

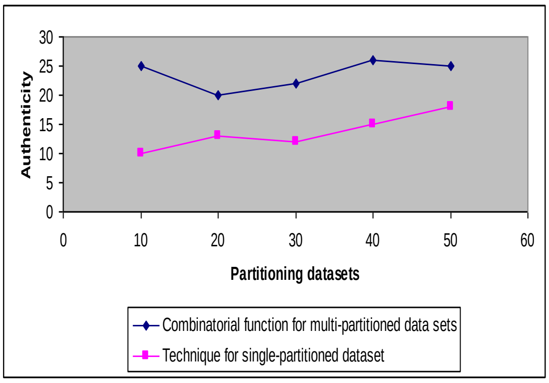
When compared to an existing technique which is to be applied for an individual partition data sets. Fig 5.1 describes the scalability of combinatorial function for multi-partitioned dataset.



5.1 Partitioning datasets vs. Scalability

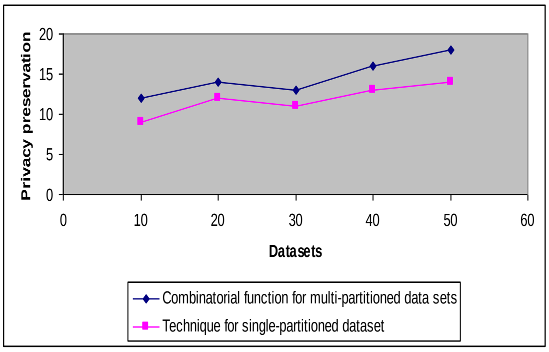
When the partitioning data set size increases, the scalability factor of the function should also be a good one to enhance the partitioning process. The proposed combinatorial function for multi-partitioned data sets provides a high scalability factor when compared to existing single-partitioned data sets.

The existing technique will support only the single partitioned data process.



5.2 Partitioning datasets vs. Authenticity

Fig 5.2 describes the authenticity of datasets.When the partitioning data set size increases, the authenticity factor of the data sets should be high to enhance the partitioning process. The proposed combinatorial function for multi-partitioned data sets provides a high authenticity when compared to existing single-partitioned data sets.



5.3 Privacy of datasets

Fig5.3 describes the Privacy of the data sets. Even the partition of data is shared among many users, the data set privacy is good in combinatorial function for multi-partitioned data sets rather than existing technique for single-partitioned data sets.

Finally, we observed that the privacy preservation combinatorial function performs efficiently with multi-partitioned data sets. The data sets has been shared efficiently with other users and it does not permit any other user to share the data anonymously.

6. CONCLUSION

In this work, we efficiently presented an effective and efficient privacy preserving combinatorial function to mine Multi-partitioned data sets. In e-business data mining models, privacy becomes a key issue. Using combinatorial function, the partitioned data sets have been shared efficiently. The experiments are conducted to estimate the scalability of the proposed combinatorial function for multi-partitioned datasets. The results showed the proposed combinatorial function for multi-partitioned datasets are scalable and allow us to determine a better secure transformation for the applications for each scenario.

7. REFERENCES

[1] Jaideep Vaidya,, Christopher W. Clifton , "Privacy-Preserving Kth Element Score over Vertically Partitioned Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 2, FEBRUARY 2009

[2] Shuguo Han, Wee Keong Ng, Li Wan, and Vincent C.S. Lee" Privacy-Preserving Gradient-Descent Methods", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 6, JUNE 2010

[3] W. Jiang and C. Clifton, "Ac-Framework for Privacy-Preserving Collaboration," Proc. Seventh SIAM Int'l Conf. Data Mining (SDM), 2007.

[4] M. Shaneck, Y. Kim, and V. Kumar, "Privacy Preserving Nearest Neighbor Search," Proc. ICDM Workshops, pp. 541-545, 2006.

[5] Dasgupta, A.; Kosara, R., "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates" Visualization and Computer Graphics, IEEE Transactions, Vol. 17, Issue: 12, Page No. 2241 – 2248, 2011.

[6] H. Yu, J. Vaidya, and X. Jiang, "Privacy-Preserving SVM Classification on Vertically Partitioned Data," Proc. 10th Pacific- Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 647- 656, Apr. 2006.

[7] D. Metzler, "Using Gradient Descent to Optimize Language Modeling Smoothing Parameters," Proc. 30th Ann. Int'l ACM SIGIR, pp. 687-688, 2007.

[8] S. Han and W.K. Ng, "Privacy-Preserving Linear Fisher Discriminant Analysis," Proc. 12th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), May 2008

[9] S. Han and W.K. Ng, "Privacy-Preserving Genetic Algorithms for Rule Discovery," Proc. Ninth Int'l Conf. Data Warehousing and Knowledge Discovery (DaWak), pp. 407-417, Sept. 2007.

[10] Rozita A. Dara, Masoud Makrehchi, Mohamed S. Kamel" Filter-Based Data Partitioning for Training Multiple Classifier Systems", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 4, APRIL 2010.

[11] Keng-Pei Lin and Ming-Syan Chen "On the Design and Analysis of the Privacy-Preserving SVM Classifier", IEEE Transactions on Knowledge and Data Engineering,Volume 23 Issue 11, November 2011

[12] Bhaduri, K., Stefanski, M.D., Srivastava, A.N., "Privacy-Preserving Outlier Detection Through Random Nonlinear Data Distortion", Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on Vol. 41 , Issue: 1, Page No. 260 – 272, 2011

[13] R. Jenssen, D. Erdogmus, K.E. Hild, J.C. Principe, and T. Eltoft, "Information Cut for Clustering Using a Gradient Descent Approach," Pattern Recognition, vol. 40, no. 3, pp. 796-806, 2007.