Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-122

Distributed Privacy Preserving Data Mining: A framework for k-anonymity based on feature set partitioning approach of vertically fragmented databases

***Abstract*— Recently, many data mining algorithms for discovering and exploiting patterns in data are developed and the amount of data about individuals that is collected and stored continues to rapidly increase. However, databases containing information about individuals may be sensitive and data mining algorithms run on such data sets may violate individual privacy. Also most organizations collect and share information for their specific needs very frequently. In such cases it is important for each organization to make sure that the privacy of the individual is not violated or sensitive information is not revealed. In this paper we have proposed a novel method to provide privacy to the data when the data is vertically partitioned and distributed over sites. In this work we presented trusted third party framework along with an application that generates k-anonymous dataset from two vertically partitioned sources without disclosing data from one site to other. K- anonymity constraint is satisfied using feature set partitioning method, which uses a genetic algorithm to search for optimal feature set partition and conventional asymmetric cryptographic technique will be used in case of trusted third party model. So in order to preserve privacy of the data trusted third party has been used and such data is first anonymized at local party using feature set partitioning method and then global classification and anonymization done at the trusted third party. We have proposed algorithm and tested different data sets for vertically partitions.**

Keywords—Distributed data mining; Privacy preserving; k-anonymity; Genetic algorithm

Jalpa Patel
Computer Engineering Department
Sarvajanik College of Engineering & Technology
Surat, India

Prof. Keyur Rana
Computer Engineering Department
Sarvajanik College of Engineering & Technology
Surat, India

I. Introduction

Large amount of information obtained from many social organization are dependent on data mining in their everyday's activities. Sharing of this kind of data is most common through various activities such as using credit cards, swapping security cards, talking over phones and use of emails. In recent years, because of the ability of computer to store great amount of data, sharing of data is very common in most organizations or even in different units of the particular organization. They use data for their specific need, for example data publishes or data mining. Also data sets typically consist of sensitive information of an individual's kind of medical and financial information. During the whole process of data mining from collection of data to discovery of knowledge these data often get exposed to several parties. As a consequence revelation of data can breach the privacy of individuals, therefore privacy is increasingly becoming an important issue and it is requires developing the algorithm for privacy preserving data mining technique and machining learning methods.

Generally data used for data mining are classified by two ways: (1) Centralized data sets (2) Distributed datasets. Furthermore, distributed data scenarios can be also categorized as horizontally partitioning data and vertically partitioning data [1]. Many papers related to privacy protection method focused on data perturbation [2] [3] [4], data encryption [5][6][7][8] and other techniques. In Privacy preserving data mining, most widely used approach is based on k-anonymity which typically aims to protect individual privacy against re-identification attack with minimal impact on the quality of the resulting data. It demands that every tuple in the dataset released be indistinguishably related to no fewer than *k* respondents [9]. In this paper we proposed a new method to achieve k-anonymity in distributed privacy preserving data mining for vertically partitioned data where data is distributed over different sites and k-anonymity constraint is satisfied by feature set partitioned approach. The main objective of our work is to preserve classification accuracy while achieving k-anonymity in vertically fragmented databases.

This paper is organized as: section II describe related work in this field, section III gives background details for k-anonymity and feature set partition approach. Following, section IV explains the proposed work with algorithm steps and experimental results analysis. Finally section V gives conclusion and future direction related to this field.

II. RELATED WORK

Many techniques have been proposed previously to preserve privacy in various data mining tasks such as classification [10], clustering and association rule mining. In this paper we mainly focus on classification task and k-anonymity model which is proposed by L. Sweeny [9] [11]. Their work uses generalization and suppression technique to achieve k-anonymity constraint. Generalization is the most common approach to achieve k-anonymity by replacing certain value with less specific value. On the other hand suppression referred to as not to release some of the values. Other methods related to that concept are bottom-up-generalization and top-down specialization (TDS) for k-anonymity have been proposed in [12] [13] respectively. More recently, Fung et al [14] presented an improved version of TDS which is called top-down refinement (TDR). Iyengar [15] has proposed novel method to search for the best set of generalization which is based on genetic algorithm. Similarly Rhonda Chaytor [16] also presented a new method using genetic algorithm to achieve *k*-anonymity. This method represents column ordering as permutations and adopts an ordered greed approach. Approaches, based on generalization has common disadvantage that it requires domain hierarchy key tree created manually. Moreover, suppression drastically reduces the quality of the data.

There are other methods developed in literature which do not use generalization and suppression techniques [17] [18]. Method given by N. Matatov et al [18] achieve k-anonymity by feature set partitioning approach in which features are decomposed into several projections and each adheres to k-anonymity constraint. More work related to feature set partitioning can be found in [19] [20] [21] and focuses on advantages over high dimensionality of data. In Feature decomposition, the goal is to decompose the original set of features into several subsets and generalizes the task of feature selection which is extensively used in data mining. Traditionally, many data mining algorithm is developed for centralized data sets. Nowadays most research focus on developing algorithm for distributed privacy preserving data mining. W. Du and Z. Zhan [22] proposed protocol for building decision tree classifier in vertically partition data between two parties. Protocol is design in such a way that no data is disclosed to other party as well as third party. Kantarcioglu and Vaidya [23] developed a privacy-preserving naive Bayes classifier for horizontally partitioned data and J. Vaidya and C. Clifton [24] presented protocol that uses naïve bayes classifier for vertically partitioned data. Both approach trained naïve bayes classifier in a way that preserve the data in horizontally partition and vertically partition.

III. BACKGROUND

As stated in section II, generalization and suppression suffers from the problem concerning data quality in data mining task. Most qualitative method has been proposed by N. Matatov et al [18] for anonymization. Their approach is known as data mining privacy by decomposition (DMPD). The main goal of this algorithm is to divide the original dataset into several disjoint projections such that each of them adheres to k-anonymity. Moreover, any attempt to rejoin the projections, results in a table that still complies with k-anonymity. This method uses a genetic algorithm to search for optimal feature set partitioning. As DMPD is limited to centralized data sets our proposed approach utilizes the advantages of that algorithm in distributed environment. This section briefly describes the concepts of k- anonymity and details of DMPD method.

A. K-anonymity

A data set provides k-anonymity if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. In k-anonymity model attributes can be categorized as confidential attribute and quasi identifiers attributes.

- *Confidential attribute:* Attribute contain private information that an individual typically does not want revealed. For example Medical record, wage, etc.
- *Quasi Identifier (QI):* A set of attributes that can be potentially linked with external information to re-identify entities. Examples of such attributes are 5-digit ZIP code, Birth date, Gender

B. Feature set partitioning approach

This approach is categorized into three main procedures. 1) Search for an optimal and valid partition based on genetic search. 2) Evaluation of a given partition 3) Combine multiple classifiers to obtain unlabeled instance classification.

- *Procedure – 1 Genetic algorithm based search*

  Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. It has become very useful in many data mining tasks and in feature selection in particular. Lior Rokach [21] has proposed an approach for feature set partitioning for classification based on genetic algorithm. DMPD use this partition presentation (Adjacency matrix encoding), crossover (Group-wise crossover) and mutation operator.

- *Procedure-2 Wrapper –based fitness evaluation*

  DMPD method uses wrapper procedure for evaluating the partitioning fitness value. The fitness value was the average accuracy over n runs, where each time n - 1 folds were used for training classifiers and one-fold for estimating the generalized accuracy of a validation dataset after combining predictions from different classifiers and obtaining final instance predictions. To obtain appropriate classifications, the wrapper uses Procedure 3 for instance classification.

- *Procedure-3 Instance Classification*

DMPD method uses naïve bayes classifier. The classification of a new instance is based on the product of the conditional probability of the target feature given the values of the input features in each subset.

IV. PROPOSED WORK

Our proposed approach ensures the privacy of sensitive data based on k-anonymity for distributed data mining where data are vertically fragmented between two parties. In this approach two parties locally anonymized their datasets using DMPD method and each anonymized dataset are sends to the trusted third party which perform global classification and anonymization by joining locally anonymized data sets. Conventional asymmetric cryptographic technique will be used in case of trusted third party model where encryption and decryption is done using RSA algorithm.

A Proposed algorithm: DPPTA

Before presenting the algorithm, we present the distributed scenario in detail considered in this research. Distributed framework considered two sites for example S1 and S2. Let T refer to an original data set, consider adult dataset [25] used for experiment. In order to get anonymized dataset quasi identifiers assumed is QI = {workclass, native-country, marital-status}. Table T is vertically partitioned between two sites S1 and S2 as T1 and T2 respectively. Where T1 consist of database DB1 = {ID, age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship} and T2 consist of database DB2 = {ID, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income}. Also, assumes S1 and S2 are semi-honest users in that they follow the execution of the algorithm but may later use the information seen to try to violate privacy.

Initially trusted third party sends mining initiation messages to both sites. Each site generates his k-anonymous data locally using DMPD method and such locally k-anonymized datasets are sends to trusted third party, encrypted by public key of trusted third party. Finally trusted third party decrypt the locally anonymized dataset with private key of its own and join the different dataset by common attribute (ID in our example) to get the global classification and anonymization. At trusted third party to get anonymized dataset DMPD algorithm is performed until it satisfies k-anonymity constraint with best valid partition. In order to make secure communication between different parties and trusted third party a conventional asymmetric cryptographic technique known as RSA is used. We call this algorithm as distributed privacy preserving trusted third party anonymizer (DPPTA). Fig.1. shows algorithm for DPPTA.

```
Inputs: Private datasets T1 and T2, Quasi identifier QI= {A1,….An} ,constraint k(anonymity level),$PU_{TTP}$ (Public key of trusted third party),$PR_{TTP}$ (Private key of trusted third party)

Output: T globally anonymized dataset
Initialize algorithm
Trusted third party sends mining message to each sites.
Site S1 and S2 runs DMPD algorithm to generate their locally k-anonymous datasets as Anon[T1]and Anon[T2]respectively.
site S1 computes E[$PU_{TTP}$, Anon[T1]],site S2 computes E[$PU_{TTP}$, Anon[T2]] and sends to trusted third party.
Trusted third party computes D[$PR_{TTP}$, Anon[T1]]and D[$PR_{TTP}$, Anon[T2]]
Trusted third party perform join operation on locally anonymization data sets, Anon[T1] ⋈ Anon[T2]
Repeat
    Perform DMPD algorithm to get globally anonymized dataset.
Until Best valid partition
Return Anon[T]
End
```

Fig. 1. Algorithm for DTTPA

B Experimental set up

The DPPTA method is evaluated in the presence of k-anonymity constraints for global classifications tasks. The comparative experiment is conducted on three different datasets such as Adult dataset, German-credit dataset and Heart dataset. All data sets are available on UCI Machine Learning Repository [25], among them adult data set is most commonly used data set for *k*-anonymity based algorithms. Adult data set consist of 48842 records and total number of attributes are 15 which is used for classification data mining task and predicts whether the income exceeds $50K/yr. Our algorithm is implemented in WEKA [26], which is a java-based environment.

D Result analysis

In this section we analyze the results obtained from experiment for our proposed method. The main goal of the analysis is

- To examine whether DPPTA satisfied the broad range of k-anonymity constraint.
- To examine effect of original classification accuracy Vs increasing level of k-anonymity constraint.

Table I summarizes the accuracy results obtained by the proposed method for different values of *k* for various datasets.

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-122

It is clearly observed form results that, there is a tradeoff between classification accuracy performance and the anonymity level.

TABLE I. ACCURACY VS. ANONYMITY-LEVEL

Data Set	*k*-anonymity level			
	1	50	100	300
Adult	83.28236	83.11032	82.63046	82.22346
German-Credit	75.12126	75.01201	74.56954	74.21234
Heart	81.51851	81.742674	80.23542	80.0121

Fig.2. plot the comparative graph for the adult dataset and illustrated that, increasing anonymity level decreases the classification accuracy.

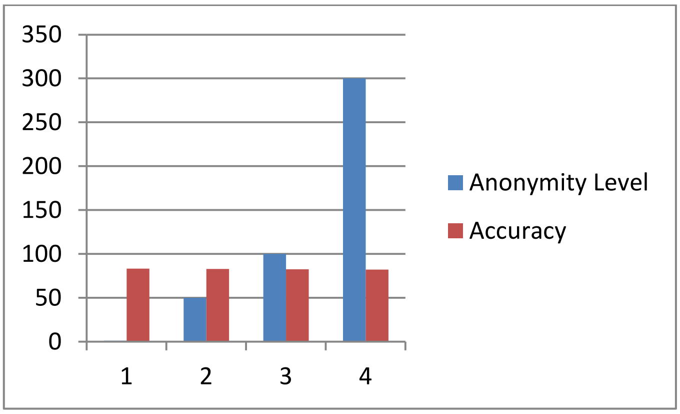

Fig. 2. Accuracy vs. anonymity level graph for adult dataset

V. CONCLUSION AND FUTURE WORK

The increasing ability to track and collect large amount of data with the use of current hardware technology has lead to an interest in the development of data mining algorithms which preserve user privacy. As per previous study, numbers of methods have recently been proposed for privacy preserving data mining in centralized data as well as distributed data environment. In this research work our proposed algorithm (DPPTA) utilizes the feature set partitioning approach (DMPD algorithm) in distributed privacy preserving data mining for vertically partitioned data. It provides more security and preserves privacy because, whole dataset is divided between two parties, so to breach privacy intruder unable to get complete dataset. Even if intruder is able to get dataset, it is anonymized at each party as well as at trusted third party who consists of complete anonymized data set. In such a way it preserves the privacy of sensitive data as well as it provide secure communication between different parties. As this algorithm is only for classification task, further we focus on other data mining tasks such as clustering and association rule mining. Also we can examine this work for other classification algorithms.

REFERENCES

[1] X. Qi , M. Zong," An Overview of Privacy Preserving Data Mining", International Conference on Environmental Science and Engineering, 2012 ,pp.1341 – 1347.

[2] L. Liu , M. Kantarcioglu, B. Thuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data", Data & Knowledge Engineering 65, 2008,pp. 5–21.

[3] R. Agrawal, R. Srikant, "Privacy-preserving data mining", In Proc. SIGMOD00, 2000, pp. 439-450.

[4] A.V. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, Privacy preserving mining of association rules, In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 217–228.

[5] B. Pinkas, "Cryptographic Techniques for Privacy Preserving Data Mining," In: SIGKDD Explorations , 2002, vol. 4, no. 2, pp. 12-19.

[6] A. Gurevich, E. Gudes, "Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation", In: IDEAS, 2006.

[7] S. Laur, H. Lipmaa, and T. Mielik¨ainen, "Cryptographically private support vector machines", In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 618–624.

[8] W. Du, Z. Zhang, "A Practical Approach to Solve Secure Multi-party Computation," In Proceedings of the 2002 workshop on New security paradigms, 2002, pp. 127-135.

[9] L. Sweeney, "*k*-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, pp. 557-570.

[10] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining", In Proc of ACM SIGMOD, 2004, Vol. 33, No. 1,pp. 50–57.

[11] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", International Journal on Uncertainty, Fuzziness and Knownledge- based Systems, 2002, pp.571–588.

[12] K. Wang, P.S. Yu, S. Chakraborty, "Bottom-up generalization: a data mining solution to privacy protection" In: Proc. of the Fourth IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, 2004, pp. 249–256.

[13] B.C.M. Fung, K. Wang, P.S. Yu, "Top-down specialization for information and privacy preservation", In: Proc. of the 21st IEEE International Conference on Data Engineering, ICDE05, IEEE Computer Society, Washington, DC, 2005, pp. 205–216.

[14] B.C.M. Fung, K. Wang, P.S. Yu, "Anonymizing classification data for privacy preservation", IEEE Transactions on Knowledge and Data Engineering, VOL. 19, NO. 5, 2007, pp. 711–725.

[15] V.S. Iyengar, Transforming data to satisfy privacy constraints, in: Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, 2002, pp. 279–288.

[16] R. Chaytor, "A Better Problem Representation for k –Anonymity", International workshopon privacy, Security, and Trust with KDD(with SIGKDD'07).

[17] Dan Zhu a, Xiao-Bai Li b, Shuning Wu, "Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining", Decision Support System 2009,pp 133-140.

[18] N. Matatov, Lior Rokach, Oded Maimon , "Privacy preserving data mining : A feature set partitioning approach." Information Science

180 ,2010, pp2696-2720

[19] O. Maimon, L. Rokach, “Improving supervised learning by feature decomposition”, in: T. Eiter, K. Schewe (Eds.), Proc. of the Second International Symposium on Foundations of Information and Knowledge Systems, Lecture Notes in Computer Science, Springer-Verlag, 2002, pp. 178–196.

[20] L. Rokach, “Decomposition methodology for classification tasks – a meta decomposer framework”, Pattern Analysis and Applications 2006, pp 257–271.

[21] L. Rokach, “Genetic algorithm-based feature set partitioning for classification problems”, Pattern Recognition 2008,pp.1693–1717

[22] W. Du and Z. Zhan, “ Building decision tree classifier on private data”. In C. Clifton and V. Estivill- astro, editors, IEEE International Conference on Data Mining workshop on Privacy, Security, and Data Mining, volume 14, 2002, pp1– 8,

[23] M. Kantarcioglu, J. Vaidya, “Privacy-Preserving Naive Bayes Classifier for Horizontally Partitioned Data”.In IEEE Workshop on Privacy-Preserving Data Mining, 2003.

[24] J. Vaidya, C. Clifton, “Privacy-Preserving Naive Bayes Classifier over vertically partitioned data”. In SIAM Conference, 2004.

[25] C.J. Merz, P.M. Murphy, UCI Repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA, 1998.

[26] I. H. Witten, E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, 2nd ed., Morgan Kaufman, San Francisco, CA, 2005.