DATA MINING IN MULTILEVEL RANDOM PERTURBATION

Mukkanti Pardhasaradhi 1*, Kamjula Ramalinga Reddy 2*

1. *M.Tech (CSE), A.M.Reddy Memorial College Of Engineering & Technology, Petulurivaripalem, Narasaraopet*
2. *Asst.Prof-CSE, A.M.Reddy Memorial College Of Engineering & Technology, Petulurivaripalem, Narasaraopet*

Abstract: *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 both to protect privileged information and to enable its use for research or other purposes. The increasing ability to track and collect large amounts 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. A recently proposed technique addresses the issue of privacy preservation by perturbing the data and reconstructing distributions at an aggregate level in order to perform the mining. This method is able to retain privacy while accessing the information implicit in the original attributes. The distribution reconstruction process naturally leads to some loss of information which is acceptable in many practical situations. Specially, we prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. We show that when a large amount of data is available, the EM algorithm provides robust estimates of the original distribution. We propose metrics for quantization and measurement of privacy preserving data mining algorithms..*

1. INTRODUCTION

We consider a scenario where two parties having private databases wish to cooperate by computing a data mining algorithm on the union of their databases. Since the databases are confidential, neither party is willing to divulge any of the contents to the other. We show how the involved data mining problem of decision tree learning can be efficiently computed, with no party learning anything other than the output itself. We demonstrate this on ID3, a well-known and influential algorithm for the task of decision tree learning. We note that extensions of ID3 are widely used in real market applications. Data mining. 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. 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. Data mining techniques are used in business and research and are becoming more and more popular with time. Confidentiality issues in data mining. 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

**** Mukkanti Pardhasaradhi***
M.Tech (CSE), A.M.Reddy Memorial College Of Engineering & Technology, Petulurivaripalem, Narasaraopet

where the sharing of data can lead to mutual gain. A key utility of large databases today is research, whether it be 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. Despite the potential gain, this is often not possible due to the confidentiality issues which arise. We address this question and show that highly efficient solutions are possible. Our scenario is the following: Let P1 and P2 be parties owning (large) private databases D1 and D2. The parties wish to apply a data mining algorithm to the joint database D1 ∪ D2 without revealing any unnecessary information about their individual databases. That is, the only information learned by P1 about D2 is that which can be learned from the output of the data mining algorithm, and vice versa. We do not assume any "trusted" third party who computes the joint output. Very large databases and efficient secure computation. We have described a model which is exactly that of multi-party computation. Therefore, there exists a secure protocol for any probabilistic polynomial-time functionality. However,, these generic solutions are very inefficient, especially when large inputs and complex algorithms are involved. Thus, in the case of private data mining, more efficient solutions are required. It is clear that any reasonable solution must have the individual parties do the majority of the

computation independently. Our solution is based on this guiding principle and, in fact, the number of bits communicated is dependent on the number of transactions by a logarithmic factor only. We remark that a necessary condition for obtaining such a private protocol is the existence of a (non-private) distributed protocol with low communication complexity. Semi-honest adversaries. In any multi-party computation setting, a malicious adversary can always alter its input. In the data mining setting, this fact can be very damaging since the adversary can define its

input to be the empty database. Then the output obtained is the result of the algorithm on the other party's database alone. Although this attack cannot be prevented, we would like to prevent a malicious party from executing any other attack. However, for this initial work we assume that the adversary is semi-honest (also known as passive). That is, it correctly follows the protocol specification, yet attempts to learn additional information by analyzing the transcript of messages received during the execution. We remark that although the semi-honest adversarial model is far weaker than the malicious model (where a party may arbitrarily deviate from the protocol specification), it is often a realistic one. This is because deviating from a specified program which may be buried in a complex application is a non-trivial task. Semi-honest adversarial behavior also models a scenario in which both parties that participate in the protocol are honest. However, following the protocol execution, an adversary may obtain a transcript of the protocol execution by breaking into one of the parties' machines.

1.1 Contributions of this paper
1.2 In this paper, we develop optimal algorithms and models based on the interesting perturbation approach proposed in

[1]. We propose a reconstruction algorithm for privacy preserving data mining, which not only converges but does so to the maximum likelihood estimate of the original distribution. This is the theoretical best that any reconstruction algorithm can achieve. This effectively means that when a large amount of data is available, the expectation maximization algorithm can reconstruct the distribution with little or almost no information loss.

We examine the problem of quantifying privacy and information loss. For example, the method in quantifies privacy without taking into account the additional knowledge that a user may obtain from the reconstructed (aggregate) distribution. We propose a privacy metric which takes into account the fact that both the perturbed individual record and the aggregate distribution are available to the user to make more accurate guesses about the possible values of the record. This privacy metric is based on the concept of mutual information between the original and perturbed records. Thus, the metrics proposed by this paper also provide a foundation for testing the effectiveness of privacy- preserving data mining algorithms in the future.

We use these proposed metrics to quantify the effects of data and perturbation parameters. Our empirical results show some simple trends of privacy-preserving data mining algorithms: (1) with increasing perturbation, the privacy level increases, but the e effectiveness of reconstruction algorithms decreases. This leads to a privacy-information loss trade O curve. (2) With increasing amount of data available, the EM-reconstruction algorithm is able to approximate the original distribution to a very high degree of precision (3) Our metrics also provides somewhat different results to those presented in [1] about the relative effectiveness of different perturbing distributions.

2 RELATED WORK

Privacy Preserving Data Mining (PPDM) was first proposed in and simultaneously. To address this problem, researchers have since proposed various solutions that fall into two broad categories based on the level of privacy protection they provide. The first category of the Secure Multiparty Computation (SMC) approach provides the strongest level of privacy; it enables mutually distrustful entities to mine their collective data without revealing anything except for what can be inferred from an entity's own input and the output of the mining operation alone,. In principle, any data mining algorithm can be implemented by using generic algorithms of SMC. However, these algorithms are extraordinarily expensive in practice, and impractical for real use. To avoid the high computational cost, various solutions that is more efficient than generic SMC algorithms have been proposed for specific mining tasks. Solutions to build decision trees over the horizontally partitioned data were proposed in. For vertically partitioned data, algorithms have been proposed to address the association rule mining, k-means clustering, and frequent pattern mining problems. The second category of the partial information hiding approach trades privacy with improved performance in the sense that malicious data miners may infer certain properties of the original data from the disguised data. Various solutions in this category allow a data owner to transform its data in different ways to hide the true values of the original data while at the same time still permit useful mining operations over the modified data. This approach can be further divided into three categories:

1) k-anonymit
2) Retention replacement
3) Data perturbation

The data perturbation approach includes two main classes of methods: additive and matrix multiplicative schemes. These methods apply mainly to continuous data. In this paper, we focus solely on the additive perturbation approach where noise is added to data values. Another relevant line of research concerns the problem of privately computing various set related operations. Two party protocols for intersection, intersection size, equijoin, and equijoin size were introduced in for honest-butcurious adversarial model. Some of the proposed protocols leak information. Similar protocols for set intersection have been proposed in. Efficient two party protocols for the private matching problem which are both secure in the malicious and honest-but-curious models were introduced in. Efficient private and threshold set intersection protocols were proposed in. While most of these protocols are equality based, algorithms in compute arbitrary join predicates leveraging the power of a secure coprocessor. Tiny trusted devices were used for secure function evaluation in.

III. Classification by Decision Tree Learning

This section briefly describes the machine learning and data mining problem of classification and ID3, a well-known algorithm for it. The presentation here is rather simplistic and very brief and we refer the reader to for an in-depth treatment of the subject. The ID3 algorithm for generating decision trees was first introduced by Quinlan in and has since become a very popular learning tool.

3.1. The Classification Problem

The aim of a classification problem is to classify transactions into one of a discrete set of possible categories. The input is a structured database comprised of attribute-value pairs. Each row of the database is a transaction and each column is an attribute taking on different values. One of the attributes in the database is designated as the class attribute; the set of possible values for this attribute being the classes. We wish to predict the class of a transaction by viewing only the non-class attributes. This can then be used to predict the class of new transactions for which the class is unknown.

For example, a bank may wish to conduct credit risk analysis in an attempt to identify non-profitable customers before giving a loan. The bank then defines "Profitable customer" (obtaining values "yes" or "no") to be the class attribute. Other database attributes may include: Home-Owner, Income, Years-of-Credit, Other-Delinquent-Accounts, and other relevant information. The bank is interested in learning rules such as If (Other-Delinquent-Accounts = 0) and (Income > 30k or Years-of-Credit > 3) then Profitable-customer = YES [accept credit-card application] A collection of such rules covering all possible transactions can then be used to classify a new customer as potentially profitable or not. The classification may also be accompanied with a probability of error. Not all classification techniques output a set of meaningful rules, we have brought just one example here. Another example application is to attempt to predict whether a woman is at high risk for an Emergency Caesarean Section, based on data gathered during the pregnancy.

There are many useful examples and it is not hard to see why this type of learning or mining task has become so popular. The success of an algorithm on a given data set is measured by the percentage of new transactions correctly classified. Although this is an important data mining (and machine learning) issue, we do not go into it here.

3.2. Decision Trees and the ID3 Algorithm

A decision tree is a rooted tree containing nodes and edges. Each internal node is a test node and corresponds to an attribute; the edges leaving a node correspond to the possible values taken on by that attribute. For example, the attribute "Home-Owner" would have two edges leaving it, one for "Yes" and one for "No." Finally, the leaves of the tree contain the expected class value for transactions matching the path from the root to that leaf. Given a decision tree, one can predict the class of a new transaction t as follows. Let the attribute of a given node v (initially the root) be A, where A obtains possible values a1, . .. , am. Then, as described, the m edges leaving v are labeled a1, . . . , am, respectively. If the value of A in t equals ai , we simply go to the son pointed to by ai .We then continue recursively until we reach a leaf. The class found in the leaf is then assigned to the transaction.

We use the following notation:

- R denotes the set of attributes.
- C denotes the class attribute.
- T denotes the set of transactions.

ID3(R, C, T)

1. If R is empty, return a leaf-node with the class value assigned to the most transactions in T.
2. If T consists of transactions which all have the same value c for the class attribute, return a leaf-node with the value c (finished classification path).
3. Otherwise:
   (a) Determine the attribute that *best* classifies the transactions in T, let it be A.
   (b) Let $a_1, \ldots, a_m$ be the values of attribute A and let $T(a_1), \ldots, T(a_m)$ be a partition of T such that every transaction in $T(a_i)$ has the attribute value a_i.
   (c) Return a tree whose root is labeled A (this is the test attribute) and has edges labeled $a_1, \ldots, a_m$ such that for every i, the edge a_i goes to the tree ID3$(R - \{A\}, C, T(a_i))$.

Fig. 1. The ID3 algorithm for decision tree learning.

The ID3 algorithm assumes that each attribute is categorical, that is containing discrete data only, in contrast to continuous data such as age, height, etc. the EM-reconstruction algorithm is able to approximate the original distribution to a very high degree of precision

IV.CONCLUSIONS AND SUMMARY

In this paper, we discussed the design and quantification of privacy-preserving data mining algorithms. We proposed an expectation-maximization algorithm which provably con- verges to the maximum-likelihood estimate of the original distribution. Thus, the algorithm provides a robust estimate of the original distribution.

We laid the foundations for quantification of privacy gain and information-loss in a theoretically accurate and method independent way. We qualified the relative effectiveness of different perturbing distributions using these metrics. Our tests also demonstrate that when the data is large then the expectation maximization algorithm can reconstruct the data distribution with almost zero information loss.

V. REFERENCES

[1] M. Ben-Or, S. Goldwasser, and A.Wigderson, Completeness theorems for noncryptographic fault tolerant distributed computation, Proceedings of the 20th Annual Symposium on the Theory of Computing (STOC), ACM, 2010, pp. 1–9.
[2] D. Boneh and M. Franklin, Efficient generation of shared RSA keys, Advances in Cryptology – CRYPTO '97, Lecture Notes in Computer Science, vol. 1233, Springer-Verlag, Berlin, 2011, pp. 425–439.
[3] R. Canetti, Security and composition of multi-party cryptographic protocols, Journal of Cryptology, vol. 13, no. 1 (2000), pp. 143–202.
[4] D. Chaum, C. Crepeau, and I. Damgard, Multiparty unconditionally secure protocols, Proceedings of the 20th Annual Symposium on the Theory of Computing (STOC), ACM, 1988, pp. 11–19.
[5] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, Private information retrieval, Proceedings of the 36th Symposium on Foundations of Computer Science (FOCS), IEEE, 1995, pp. 41–50.
[6] S. Even, O. Goldreich, and A. Lempel, A randomized protocol for signing contracts, Communications of the ACM, vol. 28 (1985), pp. 637–647.
[7] R. Fagin, M. Naor, and P. Winkler, Comparing information without leaking it, Communications of the ACM, vol. 39 (1996), pp. 77–85.
[8] J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. Strauss, and R. Wright, Secure multiparty computation of approximations, Proceedings of the 28th International Colloquium on Automata, Languages and Programming (ICALP), 2001, pp. 927–938.
[9] O. Goldreich, Secure multi-party computation. Manuscript, 1998. (Available at http://www.wisdom. weizmann.ac.il/~oded/pp.html.)
[10] O. Goldreich, S. Micali, and A.Wigderson, How to play any mental game—a completeness theorem for protocols with honest majority, Proceedings of the 19th Annual Symposium on the Theory of Computing (STOC), ACM, 1987, pp. 218–229.