Preserving Private Knowledge In Decision Tree Learning

Weiwei Fang
Information Engineering School, University of Science and Technology Beijing
Computer Center, Beijing Information Science and Technology University, China
Email: Liveinbetter@163.com

Bingru Yang and Dingli Song
Information Engineering School, University of Science and Technology Beijing
Email: sdlhr617@sohu.com

***Abstract*—Data mining over multiple data sources has become an important practical problem with applications in different areas. Although the data sources are willing to mine the union of their data, they don't want to reveal any sensitive and private information to other sources due to competition or legal concerns. In this paper, we consider two scenarios where data are vertically or horizontally partitioned over more than two parties. We focus on the classification problem, and present novel privacy preserving decision tree learning methods. Theoretical analysis and experiment results show that these methods can provide good capability of privacy preserving, accuracy and efficiency.**

***Index Terms*—Privacy Preserving, Data Mining, Decision Tree, Homomorphic encryption**

I. Introduction

In present, great advances in networking and databases technologies make it easy to distribute data across multi parties and collect data on a large scale for sharing information. Distributed data mining such as association rule mining and decision tree learning are widely used by global enterprises to obtain accurate market underlying information for their business decision. Although different enterprises are willing to collaborate with each other to data mine on the union of their data, due to legal constraints or competition among enterprises, they don't want to reveal their sensitive and private information to others during the data mining process.

There has been growing concern that use the technology of gaining knowledge from vast quantities of data is violating individual privacy. This has lead to a backlash against this technology. For example, Data-Mining Moratorium Act introduced in the U.S Senate that would have banned all data mining programs by the U.S. Department of Defense. Privacy preserving data mining (PPDM) has emerged to address this problem, and become a challenging research area in the field of data mining (DM) and knowledge discovery (KD). The main goal of preserving privacy data mining is to enable such win-win-win situations: The knowledge present in the data is extracted for use, the individual's privacy is protected, and the data holder is protected against misuse or disclosure of the data. [1].

The method of preserving privacy data mining depend on the data mining task (i.e., association rule, classification, clustering, etc.) and the data sources distribution manner (i.e., *centralize* where all transactions are stored in only one party; *horizontally* where every involving party has only a subset of transaction records, but every record contains all attributes; *vertically* where every involving party has the same numbers of transaction records, but every record contains partial attributes). In this paper, we particularly focus on applying preserving privacy data mining methods on the decision tree learning over vertically and horizontally partitioned data.

Supported by the Natural Science Foundation of China under Grant No. 69835001, Scientific Research Common Program of Beijing Municipal Commission of Education KM200811232013.

doi:10.4304/jcp.5.5.733-740

The rest of the paper is organized as follows. In the next section, we will introduce the related work in preserving privacy data mining and the contribution we did. In section 3, we provide some background technologies such as distributed decision tree learning, some secure multiparty computation and Homomorphic encryption. In section 4, we present our work of how to build a distributed decision tree over vertically partitioned data, which doesn't reveal privacy during the stages of building and classification. In section 5, we present our work of how to build a distributed decision tree over horizontally partitioned data. Section 6 shows the experimental results and privacy analysis. Conclusion is given at the last section.

II. Related work

Preserving privacy data mining provides methods that can compute or approximate the output of a data-mining algorithm without revealing at least part of the sensitive information about the data. Generally speaking, there are two approaches in privacy preserving data mining. One is using randomization techniques [2,3,4], that is, adding "noise" to the data before the data mining process, and using techniques that mitigate impact of the noise from the data mining results, however, recently there has been much debate about this kind of method, e.g., accuracy loss of mining results as altering the original data, inference problem can be derived from the reconstruction model, etc.

The other approach is using secure multiparty computation (SMC) techniques, such as secure sum, secure set union, secure size of intersection and scalar product, etc. In [6], Clifton has proposed to apply secure scalar product methods on association rules over horizontally and vertically partitioned data, respectively. In [7], Vaidya proposed algorithms on building decision tree, however, the tree on each party doesn't contain any information that belongs to other party, the drawback of this method is that the resulting class can be altered by a malicious party.

The contributions in this paper are as follows:

1) Methods proposed in this paper can be used in two contexts: vertically and horizontally partitioned data;

2) In the context of vertically partitioned data, as we apply a new classifying model, the private information can be preserved not only in the stage of building tree, but also in the classification stage;

3) In the context of horizontally partitioned data, we apply a new technology homomorphic encryption, which hasn't been used in this field; each party only can obtain the mining knowledge without leaking their own private information;

4) Both of these methods can be applied on more than two parties.

III. Background

A. Decision Tree Learning s

A decision-tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. Generally speaking, the basic algorithm for decision-tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The presentation here is rather simplistic and very brief and we refer readers to Ref. [1] for an in-depth treatment of the subject.

Obviously, the key of decision-tree induction is selecting the attribute that will best separate the samples into individual classes, for it plays an importance in not only the effectiveness of induction, but also the quality of mining rules.

B. secure multiparty computation

Secure multiparty computation (SMC) is the problem of evaluating a function of two or more parties' secret inputs, such that each party finally hold a share of the function output and no more else is revealed, except what is implied by the party's own inputs and outputs. SMC problem was firstly introduced by Yao and extend by Goldreich and others. These works use a similar methodology: the function f to be computed is represented as a Boolean circuit, and then the parties run a protocol for every gate in the circuit. Every participant gets shares of the input wires and the output wires for every gate. Since determining which share goes to which party is done randomly, a party's own share tells it nothing. Upon completion, the parties exchange their shares, enabling each to compute the final result.

In this paper, we proposed a PPDM method by applying PCIWL (Protocol for Comparing Information Without Leaking) and MNP (Mix Network Protocol), both of which belong to issues of SMC technology. We encourage readers who want deep understanding of the above two techniques to start with Ref. [8].

C. Homomorphic encryption

Computation on encrypted data does not make sense unless the encryption transformation being used has some homomorphic properties. The homomorphic encryption presented in this paper is based on a concept named privacy homomorphism [2], which was formally introduced by Rivest in 1978 as a tool for processing encrypted data. Basically, they are encryption functions E: T→T' which allow to perform a set F' of operations on ciphertext without knowledge of the decryption function D. Knowledge of D allows to recover the outcome of the corresponding set F of operations on plaintext. The security gain is especially apparent in multilevel security scenarios. That is, sensitive data will be encrypted by the classified institute, be processed by the unclassified contractor, and the result be decrypted by the classified institute [3].

Let S be a set, and S’ be a possibly different set with the same cardinality as S. Let D: S→S’ be bijective. D is decryption function, and the encryption function is E. Assign an algebraic system for plaintext operations by:

Let S be a set, and S’ be a possibly different set with the same cardinality as S. Let D: S→S’ be bijective. D is decryption function, and the encryption function is E. Assign an algebraic system for plaintext operations by:

$$U=<S;f_1,\cdots,f_k; P_1,\cdots, P_m; s_1,\cdots, s_n>$$

Where the fi is operator, the Pi is predicate, and the si is distinct constant. Assign converse computation of U with encrypted data by:

$$C=<S';f'_1,\cdots,f'_k; P'_1,\cdots, P'_m; s'_1,\cdots, s'_n>$$

Where the f’i, P’i and s’i are the encrypted version of fi, Pi and si respectively. The mapping D is called a privacy homomorphism if it satisfies the following conditions:

1)
$$\forall i(a,b,c,...)(f'_i(a,b,...)=c$$
$$\Rightarrow f_i(D(a),D(b),...)=D(c));$$

2)
$$\forall i(a,b,c,...)(P'_i(a,b,...)\equiv P_i(D(a),D(b),...));$$

3) $D(s'_i)=s_i$

In order for C and D to be of any use as a protection, the following additional constraints should be satisfied:

1) D and E are easy to compute;

2) The functions f’i and predicates P’i in C are efficiently computable;

3) E is a non-expanding cipher or an expanding cipher whose crypto text has a representation only marginally larger that the corresponding plaintext;

4) The operations and predicates in C should not be sufficient to yield an efficient computation of D.

IV. Privacy Preserving Decision Tree Learning over Vertically Partitioned Data

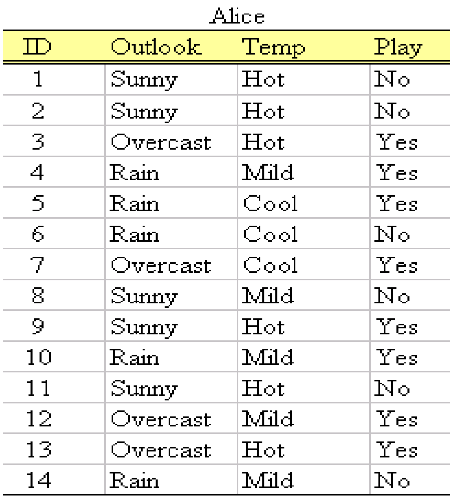
In this section we address the issue of privacy preserving distributed decision-tree mining over vertically partitioned data. Specifically, we consider a scenario in which two or more parties owning confidential databases wish to run a data-mining algorithm on the union of their databases, without revealing any original information. We propose a privacy preserving distributed decision tree learning method based on ID3 [1], which is applied in mining concentrative database and uses information entropy to choose the best prediction attribute.

Privacy preserving can mean many things [5]: Protecting specific individual values, breaking the link between values and the individual they apply to, protecting source, etc. This paper aims for a high standard of privacy: Not only individual entities are protected, but to the extent feasible even the schema (attributes and possible attribute values) are protected from disclosure.

A. Tree building

Let R be the set of condition attributes and C be the class attribute, we make assumptions that the database is vertically partitioned between k parties; each party Pi only knows its own attributes Ri, transaction ID and attribute C are known to all parties.

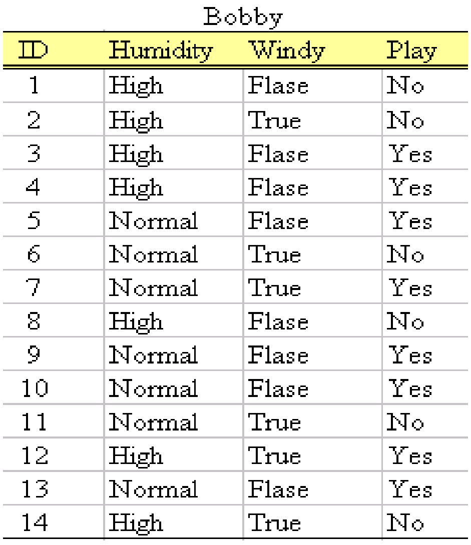
We take an example, as figure 1 shows, the class attribute is play, which is determined by four condition attributes, such as outlook, temp (possessed by Alice) and humidity, windy (possessed by Bobby).



Alice

ID	Outlook	Temp	Play
1	Sunny	Hot	No
2	Sunny	Hot	No
3	Overcast	Hot	Yes
4	Rain	Mild	Yes
5	Rain	Cool	Yes
6	Rain	Cool	No
7	Overcast	Cool	Yes
8	Sunny	Mild	No
9	Sunny	Hot	Yes
10	Rain	Mild	Yes
11	Sunny	Hot	No
12	Overcast	Mild	Yes
13	Overcast	Hot	Yes
14	Rain	Mild	No

(a) Training set in Alice



Bobby

ID	Humidity	Windy	Play
1	High	Flase	No
2	High	True	No
3	High	Flase	Yes
4	High	Flase	Yes
5	Normal	Flase	Yes
6	Normal	True	No
7	Normal	True	Yes
8	High	Flase	No
9	Normal	Flase	Yes
10	Normal	Flase	Yes
11	Normal	True	No
12	High	True	Yes
13	Normal	Flase	Yes
14	High	True	No

(b) Training set in Bobby

Figure1 Training set

If we use the traditional ID3 algorithm to mine on the union of datasets, we can obtain the public decision tree shown in figure 2, while each party's private information are all revealed.

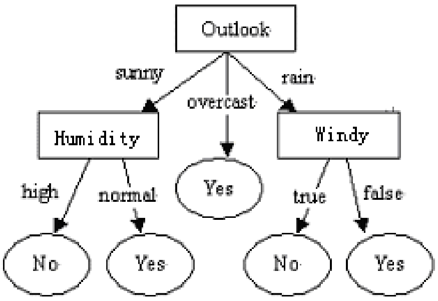


Figure 2 Public decision tree

In order to preserve each party's private data, we introduce two new notions. One is Privacy-Preserving Decision Tree, as figure 3 shows, which is stored at the miner site. The semi-honest miner only knows the basic structure of the tree, (e.g., the number of the branches at each node, the depth of each sub-tree) and which site is responsible for the decision made at each node (i.e., only know which site possesses the attribute to make decision at the node, while without the knowledge of which attribute it is and what attribute values it has); the other is Constrain Set, e.g. {AR1, BR1}, it means that this path which is form the root node to the present node (the node with the value of BR1) has determined by those attributes in the Constrain Set. When beginning to build tree, all parties will send the numbers of local attribute to miner, and the Constrain Set is initialized as {}, as Constrain Set of the present node becomes full, i.e. {AR1, BR1, AR2, BR2}, it means R is empty [1], the next node should be leaf node, which with the class attribute value c assigned to most transactions with the certain transaction IDs.

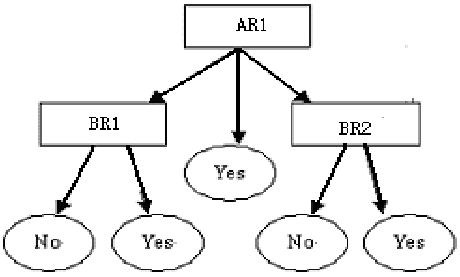


Figure 3 Privacy-preserving decision tree

Now we'll describe how it can be built and used to classify testing sets. When the miner creates a root node, it sends signal to all parties. Each party obtains the local best prediction attribute Ri by information gain measurement, then sends the attribute serial number Ri and information entropy to the miner by PCIWL (Protocol for Comparing Information Without Leaking), which ensures that no original information would be revealed at miner site or any other parties. The miner applies PCIWL to get the maximum as the global best prediction attribution, while he doesn't know the which attribute it is and what attribute values it has, he just has the knowledge that which site possesses that attribute and its' serial number, e.g., as it is shown in figure 3, the minor creates a root node AR1, which means Alice has the information at that node, and the first attribute possessed by Alice is the best prediction attribution. At the same time, the minor set {AR1} as Constraint Set of the present node.

When creating the next node, whether it's a leaf or internal node, the process is as following: Firstly, the miner sends token signal to the target party, which has possesses the best prediction attribute of previous node. Secondly, the target party receives token message, if the token signal is 0, which means Constraint Set is full, it only needs to compute the class attribute value c assigned to most transactions with the certain IDs, and send c to miner site; if the token signal is 1, which means that R isn't empty, it firstly needs to judge if all the transactions with the certain IDs have the same class attribute c, if so, then sends c to miner site; otherwise works out the intersection of transactions used previously and transactions with best prediction attribute value, and sends IDs to other parties by MNP (Mix Network Protocol), by which it guarantees that the communication process is anonymous. Thirdly, all parties compute information entropy of the local attribute corresponding to the certain IDs, and send the information entropy to the miner site by PCIWL. Finally, if the minor only receives attribute c from token party, it creates a leaf node with the value of c; if the miner receives information entropy form all parties, it chooses the maximum as the best prediction attribute, adds the attribute tag to Constant Set, and sends token to the next target party, which possesses the best prediction attribute of the present node.

B. Privacy-preserving algorithm

Assume that there are three parties named A, B and C, which respectively has ra, rb and rc condition attributes, and wants to collaboratively mining decision-tree. As the main idea we presented above, the algorithms, which comprise three parts, are as follows:

Local mining algorithm (performed by parties with token):

Input: Local training samples, token.

Output: Sending class attribute distribution to miner site, or sending IDs to other parties and information entropy to miner.

1) If token=0, computes the class attribute value c assigned to most transactions with the certain IDs, and sends c to miner site;

2) If token=1, judges if all the transactions with the certain IDs have the same class attribute c, if so, sends c to miner site;

3) If not, works out the intersection of transactions used previously and transactions with best prediction attribute value, sends IDs to other parties by MNP, and do step 4;

4) Computes information entropy and sends it to the miner site by PCIWL;

Local mining algorithm (performed by parties without token):

Input: Local training samples, transaction IDs.

Output: Sending information entropy to miner.

1) Receives transaction IDs form the token party, then computes intersection of IDs received and IDs used previously;

2) Computes information entropy corresponding to the certain IDs, and sends it to the miner site by PCIWL.

Miner site algorithm (performed by miner):

Input: Class attribute distribution from token party, or information entropy from all parties.

Output: Creating node, updating Constraint Set, sending token signal to target party.

1) If the receiving message is class attribute c from token party, creates a leaf node with the value c;

2) If the receiving message is information entropy from all parties, applies PCIWL to obtain maximum, and do step 3;

3) Creates an internal node with the value of target party's name and serial number of the best prediction attribute, adds the attribute to Constraint Set, and do step 4;

4) If Constraint Set is full, sends token=0 to the target party; otherwise sends token=1.

V. PRIVACY PRESERVING DECISION TREE LEARNING OVER HORIZONTALLY PARTITIONED DATA

In this section we address the issue of privacy preserving distributed decision-tree mining over horizontally partitioned data. Specifically, we consider a scenario in which two or more parties owning confidential databases wish to run a data-mining algorithm on the union of their databases, without revealing any original information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes.

A. Architecture

The new architecture is depicted in Figure 4, in which HE means homomorphic encryption. It's composed of the participating databases, a calculator and a miner, which uses the following basic assumptions:

1) The database is horizontally partitioned between N participants, and there is no communication between all the participants themselves;

2) The calculator only performs auxiliary computations, without knowing their meaning and having no part of the databases;

3) Although the miner manages the data mining process and reports the results to the participants, it has no part of the databases;

4) The model is semi-honest. That is, every adversary correctly follows the protocol specification, yet attempts to learn additional information by analyzing the transcript of messages received during the execution.

5) There is no external knowledge present at any side.

There are three steps when every time selecting the best predicting attribute. Firstly, every party calculates its' local Gini index, encrypts it and then transfers the encrypted data to the calculator; secondly, the calculator computes sums of certain encrypted data without knows its' meaning and transfers it to the miner; thirdly, the miner uses secret key to decrypt them, selects the best predicting attribute and broadcasts to every party.

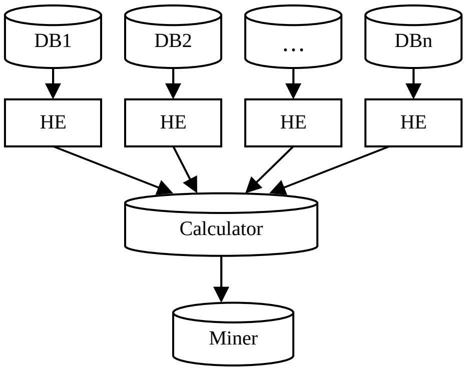


Figure4. Architecture

B. Homomorphic encryption and decryption scheme

Homomorphic encryption ensures that the computation result on two or more encrypted values is exactly the same as the encrypted result of the same computation on two or more unencrypted values.

In this paper, we proposed an additively homomorphic encryption and decryption scheme, which is as follows:

Encryption:

1) The algorithm uses a large number r, such that r= p ×q, where p and q are large security prime numbers.

2) Given x, which is a plaintext message, we compute the encrypted value y=Ep(x)=mod((x+p),r), where mod() is a common modul-operation.

Decryption:

Given y, which is a ciphertext message, we use the security key p to recover plaintext x=Dp(y)=mod(y,p).

Theorem: According to the above encryption and decryption algorithms, for every plaintext x, y and z, D (E (x)+ E (y)+ E (z)) = = x + y + z

Proof: D (E (x)+ E (y)+ E (z)) = D ((mod (x + p), n)+ (mod (y + p), n)+ (mod (z + p), n))

= D (mod (x+y+z+3p,n)) =mod (mod (x+y+z+3p,n), p)

= Mod ((mod (x + y + z), n) +mod (3p,n)), p)

= Mod ((mod (x + y + z), n), p) +mod (mod (3p,n), p)

we know n = p×q, so mod(mod(3p,n),p) =0；

As the Decryption definition says,

D(E(x)) = mod((mod(x + p),n),p) = x，

We can proof mod((mod(x + y + z),n),p) =x + y + z；

So D (E (x)+ E (y)+ E (z)) == x + y + z.

In the architecture we proposed earlier, as the calculator only sees ciphertext, has not access to the security key p, according to Domingo-Ferrer [2], with only ciphertext, it is a NP-hard problem for attacker to find the original values.

C. *Privacy-preserving algorithm*

Assume that there are three parties named A, B and C, which respectively has ma, mb and mc sample records, and wants to jointly mining decision-tree rules. As the architecture we presented above, the algorithms, which are composed of three parts, are as follows:

Local mining algorithm (performed by parties):

Input: Local training samples.

Output: Encrypted Gini Indexes of correlative attributes.

1) Scan training samples, and select sample set Sa and attribute set Ta ={Ta1, Ta2 , , , Tan} which is correlative to the present node;
2) According to definition of Gini Index [1], compute every Gini Index of attribute in set Ta, by using the formula GiniTai(Sa) = ｜Sa1｜Gini(Sa1) / ｜Sa｜+ ｜Sa2｜Gini(Sa2)/ ｜Sa｜（n≥i≥1）
3) For every element Tai (n≥i≥1)in the attribute set Ta, encrypt ma×GiniTai(Sa) (while mb×GiniTbi(Sb) for party B and mc×GiniTci(Sc) for party C) by the homomorphic encryption scheme presented above, thus get array {T’a1, T’a2 , , , T’an};
4) Send those n encrypted data to calculator;
5) Receive the best predicting attribute message Tk from miner;
6) Construct a new node’ s branches according to Tk ;

Calculate algorithm (performed by calculator):

Input: Three groups of array, which respectively contains n data.

Output: An array contained n data.

1) Receive three groups of array from three parties, which are array {T’a1, T’a2 , , , T’an}from party A, array {T’b1, T’b2 , , , T’bn} from party B and array {T’c1, T’c2 , , , T’cn } from party C;
2) Calculate each sum according to the point number in the array, T’i=T’ai+T’bi+T’ci (n≥i≥1);
3) Send an array {T’1, T’2 , , , T’n} to the miner;

Mining algorithm (performed by miner):

Input: An array {T’1, T’2 , , , T’n }.

Output: The best predicting attribute Tk.

1) Receive an array{T’1, T’2 , , , T’n}from the calculator;
2) Decrypt each T’i (n≥i≥1)by using security key p and the homomorphic decryption scheme presented above, thus get array{T1, T2 , , , Tn}, in which each element Ti(n≥i≥1)means a global computing result, that is, ma×GiniTa(Sa) + mb×GiniTb(Sb) + mc×GiniTc(Sc);
3) According to definition of Gini Index [1], select the minimum Tk from the decrypted array, which denotes the best predicting attribute;
4) Send the best predicting attribute message Tk to each party.

VI. Experiment Result

The experiment was conducted with Pentium IV3.2 GHz PC with 2GB memory on the Linux platform, and all algorithms were implemented in C/C++. We used the anonymous Web usage data of the Microsoft web site, which was created by sampling and processing the www.microsoft.com logs and donated to the Machine Learning Data Repository stored at University of California at Irvine Web Site.

We designed two sets of experiments. The first set is used to validate the effectiveness of preserving privacy algorithm on horizontally partitioned data; the second set is used to validate the effectiveness of preserving privacy algorithm on vertically partitioned data. In our experiments, we use 80% of the records as the training data and the other 20% as the testing data. We use the training data to build the decision trees, and then use the testing data to measure how accurate these trees can predict the class labels. The percentage of the correct predictions is the accuracy value in our figures. We repeat each experiment for multiple times, and each time the disguised data set is randomly generated from the same original data set.

As the resemblance of these two sets of experiments, in the following, we just show the first experiment results as an example. Figure 5 is shown that mining quality between non-privacy preserving approach and privacy-preserving approach in distributed decision-tree mining; Figure 6 is shown that privacy quality of privacy-preserving approach.

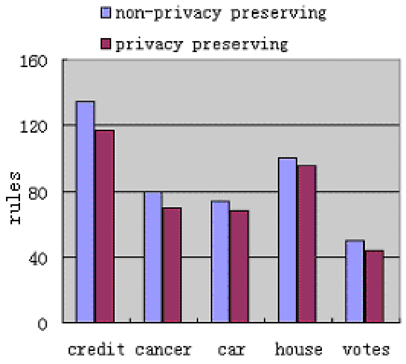


Figure5 Mining quality comparison

Figure 5 shows that compared with traditional non-privacy preserving approach, percentage of rules mined by privacy-preserving approach is at least 85%, which means although we apply privacy-preserving methods, most of the rules can also be mined; Figure 6 presents the privacy-preserving percentage is at least 82%. Experimental results show that the privacy-preserving approach we proposed can provide good capability of privacy quality without sacrificing accuracy.

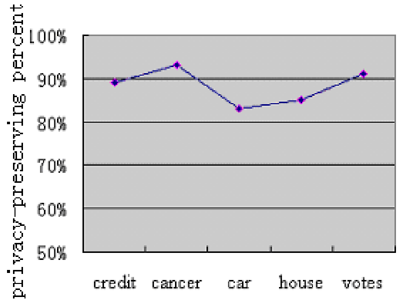


Figure 6 Privacy quality

Form the viewpoint of theoretical analysis, in the context of vertically partitioned data, when we build tree, the control is passing from site to site, except token party has the knowledge of best prediction attribute of the present node, other party even the miner doesn't know any relevant information. When we classify, the miner only knows the path of classifying process, i.e., which site handles the classifying in every step, while the information of which attribute is used to classify and values of transaction records in every party is protected. In the context of horizontally partitioned data, all row data information are encrypted, the information sent to computing center and miner are not the original information, so each party's private information are protected, and as encryption function and decryption function satisfy with Homomorphic encryption, the mining knowledge is consistent with the real results. Based on theoretical analysis and experimental results, we can conclude that methods proposed in the paper are effective.

VII. Conclusion

In this paper, we presented two privacy-preserving distributed decision-tree mining algorithms, one of which used over partitioned data is based on idea of privacy-preserving decision tree and passing control from site to site, the other of which used over horizontally partitioned data is based on the idea of homomorphic encryption. Our experimental results show that they have good capability of privacy preserving, accuracy and efficiency.

For future research, we will investigate the possibility of developing more effective and efficient algorithms. We also plan to extend our research to other tasks of data mining, like clustering and association rule, etc.

Acknowledgment

The authors wish to thank both the workmates in Data Engineering Institute, Beijing Science and Technology University and in Computer Center, Beijing Information Science and Technology University. This work was supported by the Natural Science Foundation of China under Grant No. 69835001and Scientific Research Common Program of Beijing Municipal Commission of Education KM200811232013.

References

[1] JIAWEI HAN, KAMBR M. Data mining concepts and techniques [M]. Beijing: Higher Education Press, 2001. 232 - 233.
[2] Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining [A]. In Proceedings of the 28th International Conference on Very Large Databases. Hong Kong, 2002: 682 - 693.
[3] Agrawal R, Srikant R. Privacy - preserving data mining [A]. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. United States, 2000:439 - 450.
[4] Verykios V, Bertino E. State-of-the-art in Privacy preserving Data Mining,SIGMOD, 2004,33 (1).
[5] Cliffton C, Kantarcioglu M, Vaidya J. Tools for privacy preserving distributed data mining [J]. ACM SIGKDD Explorations Newsletter ,2004 ,4 (2) :28 - 34.

[6] M.Kantarcioglous and C. Clifton. Privacy preserving distributed mining of association rules on horizontally partitioned data, ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2002), 24-31.

[7] J.Shrikant Vaidya, Privacy preserving data mining over vertically partitioned data, PH.D Thesis of Purdue University, August 2004, 28-34.

[8] Pinkas B. Cryptographic techniques for privacy-preserving data mining [J]. ACM SIGKDD Explorations Newsletter ,2006 ,4 (2) :12 - 19.

[9] Z.Yang, S.Zhong. Anonymity-preserving data collection. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, USA, August 21-24 2007

[10] S.L.Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309): 63-69, March 2004.

[11] Bingru yang,KDK Based Double-Basis Fusion Mechanism and Its Structural Model, International Journal of Artificial Intelligence Tools, 14(3), 2005： 399-423.

[12] Piatetsky-shapiro G, Matheus C J, Knowledge Discovery Work-bench for Exploring Business Databases. International Journal of Intelligent Systems, 1992,7:675-68.

[13] Yoon J P, Kerschberg L, A Frame work for Knowledge Discovery and Evolution in Databases. IEEE Transactions on Knowledge and Data Engineering, 1993,5:973-979.

[14] Yang Bing-ru, Sun Hai-hong, Xiong Fan-lun. Mining Quantitative Association Rules With Standard SQL Queries and Its Evaluation, Journal of Computer Research and Development, 39(3), 2002: 307-312.

[15] R. Agrawal, T. Imielinski, A. Swami: "Mining Associations between Sets of Items in Massive Databases", Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993, 207-216.

[16] Stephen Warshall, A Theorem on Boolean Matrices, Journal of the ACM, v.9 n.1, p.11-12, Jan. 1962.

[17] Yang Bingru, Zhou Ying. The Inner Mechanism of Knowledge Discovery System and Its Influence to KDD Mainstream Development,IC-AI'02, Las Vegas,USA.

[18] Yang Bingru.A Driving Force of Knowledge Discovery in Database Main Stream — Double Bases Cooperating Mechanism, IC-AI '02, Las Vegas, USA.

[19] Renato Coppi. A Theoretical Framework for Data Mining: the "Informational Paradigm". Computational Statistics & Data Analysis, 38(2002): 501-515.

[20] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge, CA, 2001.

[21] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, CA, 2000.

[22] Indranil Bose, Radha K. Mahapatra. Business Data Mining—A Machine Learning Perspective. Information & Management, 39 (2001): 211-225.

[23] M.S. Chen, J. Han, and P.S. Yu. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6) (1996): 866-883.

Weiwei Fang

Tutor in Beijing Information Science and Technology University; PH.D Candidate in Beijing Science and Technology University; major research field is data mining.

She has worked for morn than five years in university, mainly took lectures relevant to computer, such as programming language, database, network and etc. In 2007, she began to study in Beijing Science and Technology University as a PH.D Candidate, worked in Data Mining Institute until now. During the past two years, she has published many papers in national publication such as Computer Science, Computer Application Research, The 27th Chinese Control Conference, International Computer Science and Software Engineering Conference, etc.

Recently she has presided two projects, one is Scientific Research Common Program of Beijing Municipal Commission of Education KM200811232013, and the other is Beijing Science and Technology University Foundation 2008. She also took part in projects in Data Mining Institute, such as Natural Science Foundation of China under Grant No. 69835001, National Natural Science Foundation of China under Grant No.60875029, and etc.

Prof. Bingru Yang

He currently serves in University of Science and Technology Beijing as a chief professor, Ph.D. supervisor of School of Information Engineering and dean of Institute of Knowledge Engineering.

Prof. Dingli Song

He works as a professor in Tangshan University, teaches computer relevant lessons for more than thirty years; now he is a PH.D Candidate of University of Science and Technology Beijing, his main research field is multi-relationship data mining.