2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

Classification Rule Mining through SMC for Preserving Privacy Data Mining: A Review

Alka Gangrade [1] +, Durgesh Kumar Mishra [2], Ravindra Patel [3]

[1]Technocrats Institute of Technology, Bhopal.
[2] Acropolis Institute of Technology and Research, Indore, MP, India, 453771
[3] U.I.T., R.G.P.V., Bhopal, MP

Abstract. Data Mining and Knowledge Discovery in Databases are two new dimensions of database technology that investigate the automatic extraction for identifying hidden patterns and trends from large amount of data. Several researchers have contributed variety of algorithms for generating the classification rule by considering different cases like scalability, computation time, I/O complexity, missing attribute values, multiple decision attributes, privacy preserving of the decision system. This paper focuses on the review of the techniques for privacy preserving classification under multi-party environment. Further, the classification and secure multi-party computation algorithms have also been reviewed. The performance analysis of the algorithms has been discussed in connection with the classification.

Keywords: Classification Rule Mining, Privacy preserving, SMC.

1. Introduction

Classification Rule Mining algorithms are basically based on centralized data model that is all data is gathered into a single site. After this, running existing classification rule mining algorithms on these data. Many applications exist that are not feasible under such a methodology, leading to a need for Distributed Data Mining. Classification algorithms in Distributed Data Mining have primarily been developed from the point of view of efficiency, not security. The problem of secure distributed classification is an important one. In many situations, data is split between multiple organizations. These organizations may want to utilize all of the data to create more accurate predictive classification models while revealing neither their training data nor the instances to be classified. In many important applications, collections of mutually untrusted parties have to share information, without compromising on their privacy. In order to protect the private data, the parties perform privacy preserving computation; that is, at the end of the computation, no party knows anything except its own private data and the result.

2. Classification Rule Mining

2.1. Decision Tree Classification

Decision Tree Classifiers are used effectively in different areas: radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition etc. One of the most important features of decision tree classifiers is their ability to break down a complex decision making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret. Decision trees are representations of classifier that are easy to read and apply. Decision trees are directed graphs consisting of nodes and edges. The root node and the inner nodes represent tests and the outgoing edges represent the outcomes of the test. The leaf nodes are class labels indicated the group the example

+ Corresponding author. Tel.: + 91 9425637376; fax: +91731 4730011.
E-mail address: alkagangrade@yahoo.co.in
.

belongs to. The basic decision tree algorithm is ID3 [30] and its popular successor C4.5 [31]. C4.5 includes among others enhancements for handling continuous (non-categorical) attributes. The most important step in the tree generation process is the selection of the split attribute at each node. ID3 as well as C4.5 rely for this purpose on the information gain measure [1,6].

2.2. An Overview of Classification :

S.No.	Topic	Author	Purpose
1	SLIQ:A Fast Scalable Classifier for Data Mining	M. Mehta, *et al* [22]	Built a scalable classifier. Used pre-sorting procedure integrated with a breadth-first tree algorithm.
2	SPRINT:A Scalable Parallel Classifier for Data Mining	John Shafer, *et al* [12]	Built fast & scalable classifier that removes all the memory restrictions.
3	ScalparC : A new Scalable & Efficient Parallel Classification Algorithm for Mining Large Dataset	Mahesh V. Joshi, *et al* [18]	Scalable in both runtime and memory requirements and allowed to handle very large data sets.
4	CLOUDS:A Decision Tree Classifier for Large Datasets	Khaled Alsabti, *et al* [7]	Reduced computation time & I/O complexity also maintaining the quality of the generated tree in terms of accuracy & size.
5	RainForest : A Framework for Fast Decision Tree construction of Large Datasets	J. Gehrke, *et al* [26]	Presented a unifying framework for Decision tree classifiers that separates the scalability from the central features that determines the quality of the tree.
6	PUBLIC : A Decision Tree Classifier that Integrate Building and Pruning	R. Rastogi, *et al* [33]	Integrated the pruning phase into the building phase.

Classification has many applications in real world, such as stock planning of large superstores, medical diagnosis, etc. Some attributes may be unnecessarily transmitted to the data miner, as they are not necessary in building the classifier. This increases the risk of privacy leakage.

3. Privacy Preserving Data Mining

The objective of privacy-preserving data classification is to build accurate classifiers without disclosing private information in the data being mined.

3.1. Secure Multiparty Computation

Secure multiparty computation (SMC) enables privacy-preservation without trusted third party. SMC is one of the great achievements of modern cryptography, enabling a set of untrusting parties to compute any function of their private inputs while revealing nothing but the result of the function [11,21,25]. In order to make computation secure, for this purpose we allow non-determinism in the exact values sent in intermediate communication (like encrypt with a randomly chosen key) and show that a party only with its input and output can generate a “predicated” intermediate computation that is likely to be as actual value [1,5,9,20].

3.2. An Overview of Privacy Preserving Data Mining :

S.No.	Topic	Author	Purpose
1	Secure Multi Party Computation Problems & Their Applications: A Review & Open Problems	Wenliang Du & Mikhail J. Atallah [20]	Defined various SMC problems for their specific computations such as Privacy Preserving Data mining, Privacy Preserving Intrusion Detection, Privacy Preserving Geometric Computation.
2	Building Decision Tree Classifier on Private Data	Wenliong Du & Zhijun Zhan [14]	Built a Decision Tree Classifier on Vertically Partitioned data for preserving privacy.
3	Leveraging the “Multi” in Secure Multi-Party Computation	Jaideep Vaidya & Chris Clifton	Surveyed approaches to secure multi-party computation & gave a method where by an efficient

		[2]	protocol for two parties using an untrusted third party can be used an efficient peer-to-peer SMC protocol.
4	State-of-the-art in Privacy Preserving Data Mining	V. S. Verykios, *et al* [4]	Presented on overview of the new & rapidly emerging research area of privacy preserving data mining, also classify the techniques, review & evaluation of privacy preserving algorithms.
5	A New Scheme on Privacy Preserving Data Classification	Zan Zhang, *et al* [16]	Introduced an algebraic-technique-base scheme & compared with randomization approach.
6	Privacy Preserving Decision Tree Learning over multiple parties	F. Emekci, *et al* [3]	Focused on the classification problem & present an efficient algorithm for building a Decision Tree in a Privacy Preserving Manner using ID3 algorithm.
7	Tools for Privacy Preserving Distributed Data Mining	Chris Clifton, *et al* [5]	Presented some tools & shows how they can be used to solve several privacy preserving data mining problems.
8	Privacy Preserving Decision Tree Learning over Vertically Partitioned Data	Weiwei Fang, Bingru Yang [10]	Focused on the classification problem on vertically partitioned data more than two parties & presented a novel privacy preserving Decision Tree learning method.
9	Privacy Preserving Decision Tree over Vertically Partitioned Data	Jaideep Vaidya and Chris Clifton, *et al* [17]	Tackled the problem of classification & introduced a generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties.
10	Privacy Preserving Data Mining	R. Agrawal & R. Shrikant [11]	Proposed a novel reconstruction procedure to accurately estimate the distribution of original data value by these reconstructed distributions.
11	Privacy Preserving Data Mining	Y. Lindell, B. Pinkas [13]	Focused on the problem of Decision Tree learning with the ID3 algorithm & protocol is much more efficient.
12	Privacy Preserving Naive Bayes Classification	J. Vaidya, *et al* [15]	Presented protocols to develop a Naive Bayes classifier on both vertically and horizontally partitioned data.
13	Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned data	M. Kantarcioglu & J. Vaidya [27]	Presented protocols to develop a Naive Bayes classifier on horizontally partitioned data.
14	Privacy Preserving Naive Bayes Classifier for Vertically Partitioned data	J. Vaidya & C. Clifton [19]	Presented protocols to develop a Naive Bayes classifier on vertically partitioned data.
15	Information Sharing Across Private Data	R. Agrawal, *et al* [28]	Formalized the notion of minimal information sharing across private databases, developed protocols for intersection, equijoin, intersection size, equijoin size.
16	A Framework for high accuracy Privacy Preserving Mining	Shipra Agrawal & Jayant R. Harista [29]	Proposed a novel perturbation mechanism wherein the matrix elements are themselves characterized as random variables.
17	Using Randomized response techniques for Privacy Preserving Data Mining	Wenliang Du & Zhijun Zhan [34]	Proposed to use the randomized response techniques to conduct the data mining computation & built decision tree classifiers from the disguised data.
18	Induction of Decision Tree	J. R. Quinlan [30]	Summarized an approach to synthesizing decision trees and it describes one such system, ID3, in detail.
19	Cryptographic Techniques for Privacy Preserving Data Mining	Benny Pinkas [8]	Intended to demonstrate basic ideas from a large body of cryptographic research on secure distributed

			computation & their applications to data mining.
20	How to generate and exchange secrets	A. C. Yao [32]	Introduced a new tool for controlling the knowledge transfer process in cryptographic protocol design & applied to solve the two-party cryptographic problems.
21	Defining Privacy for Data Mining	Chris Clifton, *et al* [21]	Provides a framework and metrics for discussing the meaning of privacy preserving data mining, as a foundation for further research in this field.
22	A Framework for Privacy Preserving Classification in Data Mining	Md. Zahidul Islam and Ljiljana Brankovic [23]	Proposed a noise addition framework for protecting privacy of sensitive information used for data mining purposes. The framework does not distinguish between confidential and non-confidential attributes but rather adds noise to all of them.
23	Privacy-Preserving Decision Tree Mining Based on Random Substitutions	Jim Dowd, Shouhuai Xu, and Weining Zhang [24]	Presented a data perturbation technique based on random substitutions & showed that the resulting privacy-preserving decision tree mining method is immune to attacks that are seemingly relevant. Systematic experiments show that it is also effective.

4. Acknowledgements

We are thankful to reviewer committee of conference for excellent comments to enhance the work. We are also thankful to the University and the College for their support.

5. References

[1] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques. Indian Reprint ISBN-81-8147-049-4, *Elsevier.*

[2] Jaideep Vaidya, Chris Clifton. *Leveraging the "Multi" in Secure Multi-Party Computation.*

[3] F. Emekci , O.D. Sahin, D. Agrawal, A. El Abbadi. *Privacy preserving decision tree learning over multiple parties. Data & Knowledge Engineering* 63, 2007, pp. 348-361.

[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. *ACM SIGKDD Explorations Newsletter*, 2004, 4(2): 28-34.

[6] Arun K Pujari. Data Mining Techniques. *Universities Press(India) 13th Impression* 2007.

[7] K. Alsabti, S. Ranka, V. Singh. CLOUDS: *A Decision Tree Classifier for Large Datasets*. In Proc. KDD-98, New York City, New York, 1998.

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

[9] Andrew C. Yao. Protocols for secure computation. *In Proc. of 23rd IEEE Symposium on Foundations of Computer Science (FOCS),* 1982, pp. 160-164.

[10] Weiwei Fang, Bingru Yang. Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data. *In Proc. of the 2008 International Conference on Computer Science & Software Engineering.*

[11] R. Agrawal, R. Srikant. Privacy Preserving Data mining. *In proc. of the ACM SIGMOD on Management of data, Dallas, TX USA,* May 15-18, 2000, pp. 439-450.

[12] J. Shafer, R. Agrawal, M. Mehta. SPRINT: A Scalable Parallel Classifier for Data Mining. *In Proc. of VLDB'96, Mumbai, India, Morgan Kaufmann,* 1996, pp. 544-555.

[13] Yehuda Lindell, Benny Pinkas. Privacy preserving data mining. *Journal of Cryptology* 15(3), 2002, pp. 177-206.

[14] Wenliang Du, Zhijun Zhan. *Building decision tree classifier on private data. In CRPITS,* 2002, pp. 1-8.