Performance Analysis of Privacy Preserving Naïve Bayes Classifiers for Distributed Databases

Alka Gangrade[1] and Ravindra Patel[2]

[1] **Technocrats Institute of Technology**
Bhopal, MP, India

[2] **Deptt. of MCA, U.I.T., R.G.P.V.**
Bhopal, MP, India

Abstract

The problem of secure and fast distributed classification is an important one. The main focus of the paper is on privacy preserving distributed classification rule mining. This research paper addresses the performance analysis of privacy preserving Naïve Bayes classifiers for horizontal and vertical partitioned databases. The Naïve Bayes classifier is a simple but efficient baseline classifier. We compare the performance of our two proposed privacy preserving Naïve Bayes protocols with basic Naïve Bayes classifier (NBC). First protocol used Un-trusted Third Party (UTP) for privacy preserving Naïve Bayes classifier for horizontally partitioned data and second protocol used secure multiplication protocol for privacy preserving Naïve Bayes classifier for vertically partitioned data. The results analysis shows that our protocols execution time is less than the existing NBC execution time since in our protocol, all parties individually calculate their probability or model parameters as an intermediate result and transfer only these intermediate results for further calculations. Accuracy of test data is same because calculated model parameters of training data are same. Our protocols are very easy to follow, understand with minimum efforts, secure and fast.

Keywords: *Privacy preserving, horizontally partitioned, vertically partitioned, SMC, UTP, Naïve Bayes.*

1. Introduction

In recent times, there have been growing interests on how to preserve the privacy in data mining when sources of data are distributed across multi-parties. Extractions of useful knowledge from huge amount of data need different techniques and strategies. These techniques are preferred to be faster, more accurate and above all very intelligent. Privacy preserving data mining is one of the most demanding research areas within the data mining community. In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. Nowadays privacy preserving data mining is the most challenging research area within the data mining society. In several cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end [1]. This requires secure protocols for sharing the information across the different parties. The data may be distributed in two ways: Horizontal partitioned data and Vertical partitioned data. Horizontal partition means, where different sites have different sets of records containing the same attributes. Vertical partition means, where different sites have different attributes of the same sets of records [2]. The classification is very important step in data mining for interpretation of useful information. In this paper, we analyze the performance of our two privacy preserving Naïve Bayes classifiers for distributed databases.

1.1 Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem [3]. Studies comparing classification algorithm have found a simple Bayesian classifier known as the Naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifier. Bayesian classifiers have also exhibited high accuracy and speed when applied to large database. In order to see how a privacy preserving Naïve Bayesian classifier is constructed, we need to address two issues: how to calculate the probability for each attribute and how to classify a new tuple [4, 5, 6].

1.2 Organization of the paper

This paper compares the performance of privacy preserving SMC protocols based on Naïve Bayes classification rule mining. The rest of the paper is organized as follows: In Section 2, we discuss the related work. Section 3, describes experimental result analysis of our 3-Layer Privacy Preserving Horizontally Partitioned NBC (3LPPHPNBC). Section 4, describes experimental result analysis of 3-Layer Privacy Preserving Vertically Partitioned NBC (3LPPVPNBC). Section 5, we conclude our paper with the discussion of the future work.

2. Related Work

Privacy preserving data mining has been a dynamic research area for a decade. A lot of work is going on by the researcher on privacy preserving classification in distributed data mining. Yao described the first SMC problem [7]. SMC allows parties with similar background to compute result upon their private data, minimizing the threat of disclosure was explained [8]. A variety of tools discussed and how they can be used to solve several privacy preserving data mining problem [9]. We now give some of the related work in this area. Preserving customer privacy by distorting the data values proposed by Agrawal and Srikant [10]. There have been several approaches to support privacy preserving data mining over multi-party without using third parties [10, 11]. Since then, there has been work improving this approach in various ways [12]. Cryptographic research on secure distributed computation and their applications to data mining were demonstrated by Pinkas Benny [1].

Classification is one of the most widespread data mining problems come across in real life. Common classification techniques have been widely studied for over two decades. The classifier is usually represented by decision trees, Naïve Bayes classification and neural networks. Quinlan proposed first ID3 decision tree classification algorithm in [13]. A secure algorithm proposed to build a decision tree using ID3 over horizontally partitioned data between two parties using SMC [14]. An innovative privacy preserving distributed decision tree learning algorithm [15] that is based on [16]. The ID3 algorithm is scalable in terms of computation and communication cost, and therefore it can be run even when there is a large number of parties involved and eliminate the need for third party and propose a new method without using third parties. A generalized privacy preserving variant of the ID3 algorithm for horizontally partitioned data distributed over multi parties introduced in [17, 18] and for vertically partitioned data distributed over two or more parties introduced in [19, 20, 21, 22] and. Privacy preserving Naïve Bayes classification for horizontally partitioned data introduced in [4, 23] and vertically partitioned data introduced in [5, 6, 24].

3. Performance Analysis of 3LPPHPNBC

3.1 Introduction

Our three layer protocol uses an Un-trusted Third Party (UTP). We have already studied how to calculate the model parameters for privacy preserving NBC, where database is horizontally partitioned and communicate their intermediate results to the UTP not their private data. For system architecture and details of Algorithms/Protocols refer [23]. The protocol presented is very efficient. However, they compromise a little on security. At the end of the protocol, all parties learn only model parameters not the attribute value. We present a protocol which does not reveal anything except the final classifier or model parameters. An UTP allows well-designed solutions that meet privacy constraints and achieve acceptable performance.

3.2 Security Analysis

We analyzed the security of all the algorithms layer wise. In our system architecture there are three layers. Input layer and output layer computations are done by the individual party by their own. Input layer transfer only intermediate results to the UTP, then UTP apply some computation on these results and send the model parameters or probabilities to all party. Thus there is no privacy leakage. After that each party is able to classify the new tuple. Privacy of all party is maintained.

3.3 Experiment and Results

The experiment was conducted with i3 II generation processor with 2GB RAM having 500GB hard disk. For implementing of algorithms, we use the software Net-Beans IDE (version 6.9). It is an open-source integrated development environment. Net-Beans IDE supports development of all Java applications and integrated these algorithms into Weka version 3.6. Weka is a data mining tool that is used to perform various data mining algorithms. It consists of various clustering algorithms, classification algorithms and a number of tools to evaluate the data mining algorithm performance. Here we have integrated the API of Weka into java such that by using various functions of the Weka we can develop various applications. It is software through which we can analyze the data on different datasets [25].

In this section we show the details of our implementation of 3-Layer Privacy Preserving Horizontally Partitioned NBC (3LPPHPNBC) algorithm on modern hardware and

IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 3, March 2013
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org

show the experimental results on Student Education datasets shown by Table 1.

Table 1: Student Education Dataset

Attribute Name	*No. of values*	*Category*
Age	3	<=30, 31..40, >40
Income	3	High, Medium, Low
Technical	3	Best, Better, Good
Student	2	Yes, No
Credit_rating	2	Fair, Excellent
Buy_computer (Class)	2	Yes, No

In our experiments, we showed how collaboration reduces time for calculating model parameters. We also show that accuracy of our algorithm on test data is almost same with compare to basic NBC. Here we use 50% of the datasets for training and use 50% for measuring the accuracy of classification. Here we are discussing two-party and multi-party cases.

3.3.1 Two-party Case

Table 2: Execution time comparison for calculating model parameters

Number of Tuples	*Size of Dataset*	*NBC Execution Time (ms)*	*3LPPHPNBC Execution Time (ms)*
10500	277KB	372	158
21000	555KB	750	246
42000	1.08MB	1226	438
84000	2.16MB	2378	879
100000	2.58MB	2819	1088

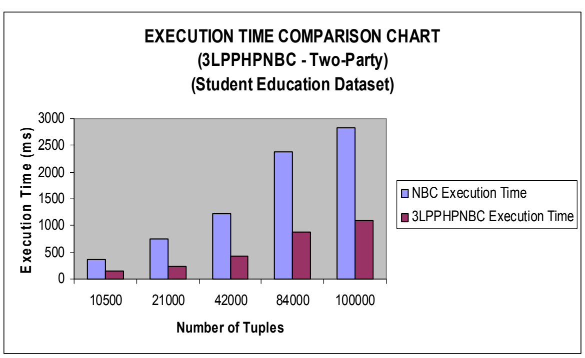

Fig. 1 Execution time comparison chart (NBC Vs 3LPPHPNBC)

Table 2 and Fig. 1 show the comparative analysis of execution time of the existing NBC and the proposed 3-layer privacy preserving horizontal partitioned NBC. It is found that our proposed algorithm takes less time for calculating model parameters.

Table 3: Comparison of correctly classified test data (Accuracy)

Number of Tuples	*NBC Accuracy (%)*	*3LPPHPNBC Accuracy (%)*
10500	75.338 %	75.3333 %
21000	75.338 %	75.338 %
42000	75.338 %	75.338 %
84000	75.338 %	75.338 %
100000	75.34 %	75.338 %

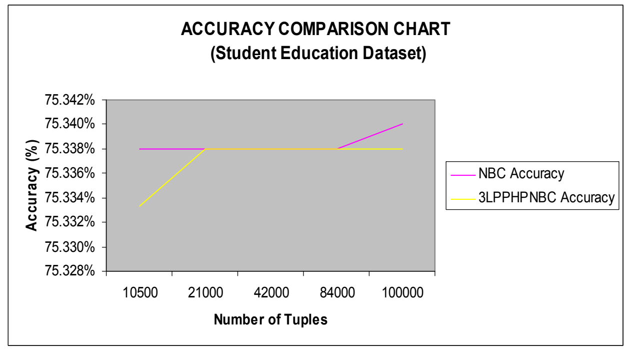

Fig. 2 Accuracy comparison chart (NBC Vs 3LPPHPNBC)

Table 3 and Fig. 2 show the accuracy comparison for classifying the test data. We found that accuracy of existing NBC and the proposed 3LPPHPNBC is same because model parameters calculated by both the algorithm is same.

3.3.2 Multi-party Case

Table 4 and Fig. 3 show the comparison of time to calculate the model parameters while using multi-party. Here time decreases if the number of parties involved increases. There is no change in accuracy.

Table 4: Execution Time comparison using Multi-party

No. of Tuples	*3LPPHPNBC Two Party Execution Time (ms)*	*3LPPHPNBC Three Party Execution Time (ms)*	*3LPPHPNBC Four Party Execution Time (ms)*
10500	158	144	133
21000	246	232	215
42000	438	408	384
84000	879	821	795
100000	1088	1044	949

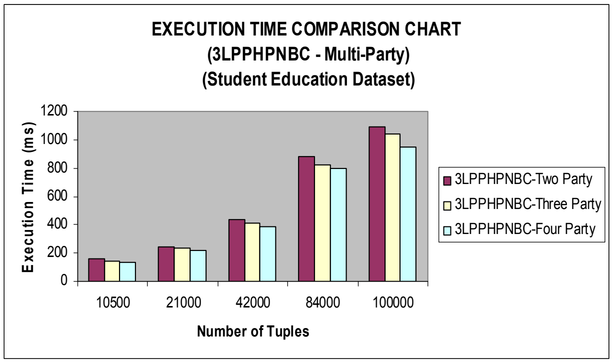

Fig. 3 Execution time comparison chart (Multi-party)

4. Performance Analysis of 3LPPVPNBC

4.1 Introduction

In this section, we focus on performance analysis of privacy preserving Naïve Bayes classification in distributed environment using secure multiplication protocol where data are vertically partitioned. We developed new and simple algorithm to classify the vertically partitioned data. The main advantage of our work over the existing one is that each party cannot gather the other's private data and it is simple and its performance is unmatched by any previous algorithm. Every party separately calculates model parameters or probability for each and every attribute then calculates total probability of each class using secure multiplication protocol. First party drives the protocol and finally finds out the class having maximum probability to classify the new tuple. Building the classifier model for vertically partitioned data, each party has complete information about the attributes present with them. Each party can locally compute the model parameters of the attributes. Security is needed only when the party classifying the new tuple. The system architecture and details of Algorithms/Protocol i.e. procedure for calculating the model parameters for nominal attributes and classifying new tuple are described in [24]. With our algorithms, the execution time required for calculating the model parameters is reduced compared to existing algorithms and the accuracy of test data set is almost same.

4.2 Security Analysis

We analyze the security of all the algorithms layer wise. In over system architecture there are three layers. We initially analyze the security of the primary algorithms, then the security of the complete algorithm. Some of the primary algorithms are executed by the party itself so there is no question of privacy leakage. Input layer computations are done by the individual party their own. Master or driving party is used secure multiplication protocol for calculating total probabilities to classify the new tuple at intermediate layer and find the class having maximum probability at output layer. Protocol secured the information transfer by other parties, thus overall privacy is maintained. Thus there is no privacy leakage.

4.3 Experiment and Results

The experiments are conducted on the same H/W and S/W system mentioned in previous section (Refer Section 3). First we apply 3-layer privacy preserving vertically partitioned NBC (3LPPVPNBC) algorithm on Student Education dataset (Refer Table 1) on two party and then multi-party case.

4.3.1 Two-party Case

Table 5: Execution time comparison for calculating model parameters

Number of Tuples	*Size of Dataset*	*NBC Execution Time (ms)*	*3LPPVPNBC Execution Time (ms)*
10500	277KB	372	145
21000	555KB	750	221
42000	1.08MB	1226	395
84000	2.16MB	2378	797
100000	2.58MB	2819	938

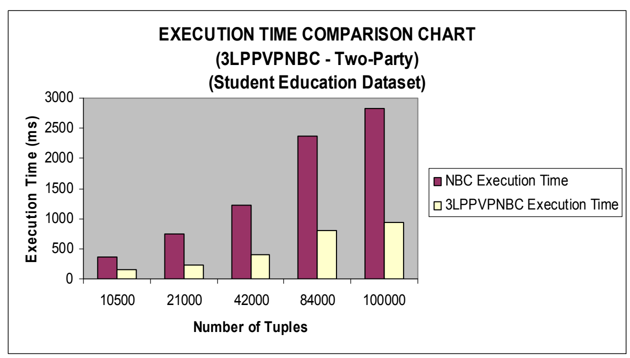

Fig. 4 Execution time comparison chart (NBC Vs 3LPPVPNBC)

Table 5 and Fig. 4 show the comparative analysis of execution time of the existing NBC and the proposed 3-layer privacy preserving vertical partitioned NBC. It is found that our proposed algorithm takes less time for calculating model parameters.

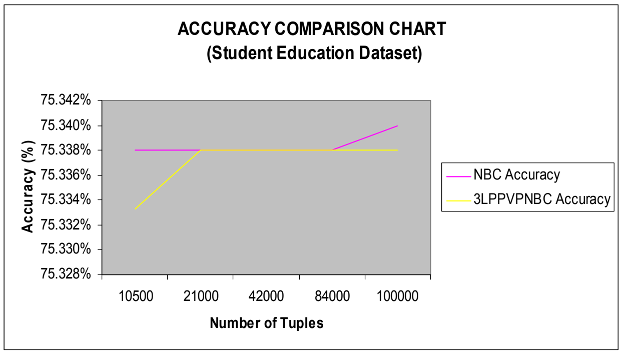

Fig. 5 Accuracy comparison chart (NBC Vs 3LPPVPNBC)

Fig. 5 shows the accuracy comparison for classifying the test data. We found that accuracy of existing NBC and the proposed 3LPPVPNBC is same because model parameters calculated by both the algorithm is same.

4.3.2 Multi-party Case

Fig. 6 shows the comparison of time for calculating model parameters while using multi-party. It is found that time decreases if the number of parties involved increases and accuracy is almost same.

Table 6: Execution Time comparison using Multi-party

No. of Tuples	*3LPPVPNBC Two Party Execution Time (ms)*	*3LPPVPNBC Three Party Execution Time (ms)*	*3LPPVPNBC Four Party Execution Time (ms)*
10500	145	122	115
21000	221	201	191
42000	395	373	361
84000	797	773	760
100000	938	908	892

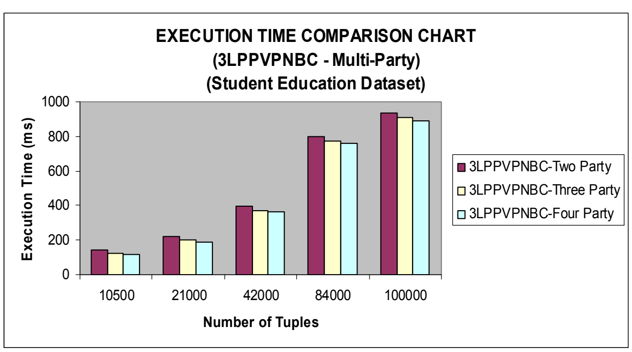

Fig. 6 Execution time comparison chart (Multi-party)

5. Conclusions and Future Work

In this paper we have presented performance of our first protocol which uses Un-trusted Third Party (UTP) for horizontally partitioned databases. Here all party transfer their counts of each class for every attribute value as an intermediate results form to UTP only not the original data and UTP calculates total probability. Through the communication between UTP and all party, final result is carried out. It requires less memory space and provides fast and easy calculations. Using this protocol, data will almost secure and privacy of individual will be maintained. We have also presented performance of our privacy preserving NBC algorithm which uses secure multiplication protocol for vertically partitioned databases. Here all party calculates the probability of each class value for every attribute value by their own. The first party drives the protocol. First party calculates the total probability and finds the class having highest probability of new tuple while maintaining privacy of all participating parties.

According to our experiments, our 3LPPVPNBC algorithm calculates model parameters faster than the 3LPPHPNBC since in 3LPPVPNBC, all party calculates model parameters by their own. While our 3LPPVPNBC classify new tuple slower than the 3LPPHPNBC because vertical partition database needs the collaboration of all party to classify the new tuple.

In our protocols, all parties are involved for calculation of model parameters. However, there can be other such mechanism needs to be addressed which minimize the involvement of parties for distributed classification rule mining.

We have addressed distributed privacy preserving classification rule mining methods where data are distributed either horizontally or vertically. However, there can be other mechanism needs to be addressed where data are distributed horizontally as well as vertically both i.e. grid partitioned.

Further development of the protocol is expected for joining multi-party using Trusted Third Party (TTP). The major challenges in privacy preserving classification rule mining is to maintain the security of all participating party, minimize the execution time to build the model of training data and improve the accuracy of the test data.

Acknowledgments

We are grateful to the University and the College for their support. We express gratitude to my colleagues for their technical support and the referees for their beneficial suggestions.

References

[1] B. Pinkas, “Cryptographic techniques for privacy-preserving data mining”, ACM SIGKDD Explorations Newsletter, Vol. 4, No. 2, 2006, pp. 12-19.
[2] C. C. Aggarwal, and P. S. Yu., Privacy-Preserving Data Mining: Models and Algorithms (London, Kluwer Academic Publishers Boston).
[3] J. Han, and M. Kamber, Data Mining: Concepts and Techniques (India, Elsevier).
[4] M. Kantarcioglu, and J. Vaidya, “Privacy preserving naive Bayes classifier for horizontally partitioned data”, IEEE ICDM Workshop on Privacy Preserving Data Mining, Melbourne, FL, November 2003, pp. 3-9.
[5] J. Vaidya, and C. Clifton, “Privacy preserving naive Bayes classifier on vertically partitioned data”, SIAM International Conference on Data Mining, Lake Buena Vista, Florida, April 2004, pp. 22-24.
[6] Z. Yang, and R. Wright, “Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data”, IEEE Transactions on Data Knowledge Engineering, Vol. 18, No. 9, April 2006, pp. 1253-1264.
[7] A. C. Yao, “Protocols for secure computation”, 23rd IEEE Symposium on Foundations of Computer Science (FOCS), 1982, pp. 160-164.
[8] W. Du, and Mikhail J. Attalah, “Secure multi-problem computation problems and their applications: A review and open problems”, Tech. Report CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN 47906, 2001.
[9] C. Cliffton, M. Kantarcioglu, and J. Vaidya, “Tools for privacy preserving distributed data mining”, ACM SIGKDD Explorations Newsletter, Vol. 4, No. 2, 2004, pp. 28-34.
[10] R. Agrawal, and R. Srikant, “Privacy preserving data mining”, ACM SIGMOD on Management of data, Dallas, TX USA, May 15-18, 2000, pp. 439-450.
[11] V. Verykios, and E. Bertino, “State-of-the-art in Privacy preserving Data Mining”, SIGMOD Record, Vol. 33, No. 1, 2004, pp. 50-57.
[12] D. Agrawal, and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms”, Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 21-23 2001, pp. 247–255.
[13] J. R. Quinlan, “Induction of decision trees”, Jude W. Shavlik, Thomas G. Dietterich, (Eds.), Readings in Machine Learning, Morgan Kaufmann, 1, 1990, pp. 81–106.
[14] Y. Lindell, and B. Pinkas, “Privacy preserving data mining”, Journal of Cryptology, Vol. 15, No. 3, 2002, pp. 177–206.
[15] F. Emekci, O. D. Sahin, D. Agrawal, and A. El Abbadi, “Privacy preserving decision tree learning over multiple parties”, Data and Knowledge Engineering Vol. 63, 2007, pp. 348-361.
[16] A. Shamir, “How to share a secret”, Communications of the ACM, Vol. 22, No. 11, 1979, pp. 612-613.
[17] A. Gangrade, and R. Patel, “A novel protocol for privacy preserving decision tree over horizontally partitioned data”, International Journal of Advanced Research in Computer Science, Vol. 2, No. 1, 2011, pp. 305-309.
[18] A. Gangrade, and R. Patel, “Privacy Preserving Two-Layer Decision Tree Classifier for Multiparty Databases”, International Journal of Computer and Information Technology (2277 – 0764), Vol. 1, No. 1, 2012, pp. 77-82.
[19] W. Du, and Z. Zhan, “Building decision tree classifier on private data”, In CRPITS, 2002, pp. 1–8.
[20] J. Vaidya, C. Clifton, M. Kantarcioglu, and A. S. Patterson, “Privacy-preserving decision trees over vertically partitioned data”, 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, 2008, pp. 139–152.
[21] J. Shrikant Vaidya, “Privacy preserving data mining over vertically partitioned data”, doctoral diss., Purdue University, August 2004.
[22] W. Fang, and B. Yang, “Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data”, 2008 International Conference on Computer Science and Software Engineering, 2008, pp. 1049-1052.
[23] A. Gangrade, and R. Patel, “Privacy Preserving Naïve Bayes Classifier for Horizontally Distribution Scenario using Un-trusted Third Party”, IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727, Volume 7, No. 6, Nov.-Dec. 2012, pp. 04-12.
[24] A. Gangrade, and R. Patel, “Privacy Preserving Three-Layer Naïve Bayes Classifier for Vertically Partitioned Databases”, paper communicated in Journal of Information and Computing Science (JIC) ISSN: 1746-7659.
[25] I. H. Witten, E. Frank and M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques (Burlington, MA: Morgan Kaufmann, 2011).

Alka Gangrade, Ph. D. student of R.G.P.V., Bhopal (MP), India. Her research interests include privacy preserving classification rule mining in a secure manner using multi-party computation protocols.

Dr. Ravindra Patel, Reader and Head, Deptt. of Computer Applications at R.G.P.V, Bhopal (MP), India. He has been awarded for Ph. D. degree in Computer Science. He poses more than 10 years of experience in post graduate classes. He has published more

than 25 papers in International and National Journals and Conferences proceedings. He is a member of International Association of Computer Science and Information Technology (IACSIT).