An Efficient Solution for Privacy-preserving Naïve Bayes Classification in Fully Distributed Data Model

Vu Duy Hien, Luong The Dung, Hoang Duc Tho

***Abstract*—Recently, privacy preservation has become one of the most important problems in data mining and machine learning. In this paper, we propose a novel privacy-preserving Naïve Bayes classifier for the fully distributed data scenario where each record is only kept by a unique owner. Our proposed solution is based on a secure multi-party computation protocol, so that it has the capability to securely protect each data owner's privacy, as well as accurately guarantee the classification model. Furthermore, our experimental results show that the new solution is efficient enough for practical applications.**

***Tóm tắt*—Gần đây, bảo vệ tính riêng tư đã trở thành một trong những vấn đề quan trọng nhất trong khai phá dữ liệu và học máy. Trong bài báo này, chúng tôi đề xuất một bộ phân lớp Naïve Bayes đảm bảo tính riêng tư mới cho kịch bản dữ liệu phân tán đầy đủ trong đó mỗi bản ghi chỉ được giữ bởi một người sở hữu duy nhất. Giải pháp nhóm tác giả đề xuất được dựa trên tính toán bảo mật nhiều thành viên nên nó có khả năng bảo vệ an toàn sự riêng tư của mỗi người sở hữu dữ liệu cũng như đảm bảo tính chính xác của mô hình phân lớp. Hơn thế nữa, các kết quả thực nghiệm của chúng tôi chỉ ra rằng giải pháp mới đủ hiệu quả trong các ứng dụng thực tế.**

Keywords—privacy-preserving data mining and machine learning; secure multi-party computation; Naïve Bayes classification; Homomorphic encryption; Data privacy.

Từ khóa—khai phá dữ liệu và học máy đảm bảo tính riêng tư; tính toán bảo mật nhiều thành viên; phân lớp Naïve Bayes; mã hóa đồng cấu; tính riêng tư của dữ liệu.

I. INTRODUCTION

In recent years, the growth of data mining and machine learning (DM and ML) has brought many benefits to organizations and individuals. However, the processes of DM and ML can violate sensitive/private information in datasets. Hence, privacy preservation has become one of the most important issues in DM and ML fields.

In general, privacy-preserving DM and ML solutions have three important properties, i.e., privacy, accuracy, efficiency [1]. They are based on the following approaches:

Randomization approach: the original input data of privacy-preserving DM and ML solutions following this way often has randomly transformed or added noise. As a result, such solutions' performance is high, but they have a trade-off between privacy and accuracy.

Cryptography approach: such privacy-preserving DM and ML techniques are often based on secure multi-party computation protocols (SMC) using homomorphic cryptosystems. As a result, cryptography-based privacy preserving DM and ML solutions can preserve each data holder's privacy, as well as guarantee the accuracy property. However, their performance is quite poor.

Hybrid approach: privacy-preserving DM and ML methods following the hybrid approach use SMC protocols combined with randomization techniques. Hence, such solutions must balance the accuracy, privacy and efficiency properties.

In this paper, we focus on privacy-preservation solutions for Naïve Bayes algorithm, one of the most common machine learning techniques. Particularly, we investigate privacy-preserving Naïve Bayes classification (PPNBC) solutions in the fully distributed setting which is a special case of the horizontally distributed data model. In this scenario, each data record is kept by a unique holder.

Up to now, researchers have proposed many privacy-preserving Naïve Bayes classifiers for the fully and horizontally distributed data settings, and such solutions are often based on cryptography and hybrid approaches.

(i) Cryptography-based PPNBC solutions:

In 2003, Kantarcıoˇglu and Vaidya [2] first introduced a privacy-preserving Naïve Bayes classifier for the horizontally distributed data setting (a similar version can be found in [3]). The solution [2] is based on the simple sum computation protocol [4], [5] that is insecure, if there exist several colluding parties. Hence, each data holder's privacy in [2] is not securely protected.

Yang et al. [6] proposed a PPNBC solution for the fully distributed setting by executing the privacy-preserving frequency mining protocol multiple times. Consequently, although this proposal can preserve the parties' privacy, its cost is quite expensive.

In 2008, Yi et al. [7] presented a privacy-preserving Naïve Bayes classifier on distributed data using two semi-trusted mixers who do not collude together. This leads the performance of [7] to be high, but each data holder's privacy cannot be protected.

Skarkala et al. [8] proposed privacy-preserving Naive Bayes classification techniques based on the multicandidate election schema. By using the Paillier cryptosystem [9] and authentication methods, data providers are protected. Unfortunately, Skarkala et al.'s solution require each data provider to share the frequencies of his/her dataset for the miner. Thus, the private property of [8] cannot be ensured.

(ii) Hybrid approach-based PPNBC solutions:

Based on Gentry's scheme [10], Li et al. propounded privacy-preserving outsourced PPNBC [11] in the cloud model. In the training phase of [11], the evaluator approximately computes a classification model from the encrypted training set. Consequently, this solution cannot guarantee the accuracy property. Furthermore, the performance of [11] is expensive, because the Gentry's cryptosystem is costly.

Huai et al. [12] described a PPNBC solution based on the privacy-preserving aggregation protocol combined with a data perturbation technique. Moreover, Huai et al. use a trusted dealer to generate the necessary secret parameters. Thus, this solution must have a trade-off between the privacy and accuracy properties, and its computational cost is pricey.

Based on differential privacy methods and homomorphic cryptosystems, privacy-preserving Naïve Bayes classifiers [13], [14] can protect data providers' privacy. However, these solutions must have a trade-off between the data providers' privacy and the classification model's accuracy. Additionally, data providers are required to spend high costs performing the tasks.

It can be seen that the existing PPNBC solutions for the fully and horizontally distributed data settings have many disadvantages. Therefore, it is significant to construct an efficient privacy-preserving Naïve Bayes classifier that has high security level, as well as guarantees the accuracy property.

II. PRELIMINARIES

A. PROBLEM OF NAIVE BAYES CLASSIFICATION IN THE FULLY DISTRIBUTED SETTING WITH PRIVACY CONSTRAINTS

Assume that there are a dataset D consisting of m independent attributes $\{A_1, \dots, A_m\}$ and one class label attribute, in which each attribute A_j has a defined set of l_j values $\{a_j^1, \dots, a_j^{l_j}\}$ $(j = \overline{1,m})$, and the set of class labels is $\{L_1, \dots, L_t\}$. The dataset D has n vectors $\{u_1, \dots, u_n\}$ of $(m + 1)$ elements, where each data vector u_i is privately kept by the data owner U_i $(i = \overline{1,n})$.

To build the Naive Bayes classification model based on the dataset D with privacy constraints, a miner needs to compute the necessary probabilities using the following frequency values:

Number of data vectors that their class label is L_k $(k = \overline{1,t})$, denoted as $\#(L_k)$.

Number of data vectors $\#(a_j^r, L_k)$ that their j^{th} attribute is a_j^r, and their class label is L_k ($j = \overline{1,m};\ r = \overline{1,l_j}\ ; k = \overline{1,t}$).

while each data owner discloses nothing about his/her data vector.

Basically, these frequency values are often privately calculated by using secure sum computation or privacy-preserving frequency mining protocols. This paper nominates the secure multi-party sum protocol [15] as one of the most suitable and efficient candidates for privately computing the above frequency values used in Naïve Bayes classifier. In the other words, by executing the secure multi-party sum protocol [15] multiple times, we obtain a privacy-preserving Naïve Bayes classifier in the semi-honest model for the fully distributed data setting.

B. An efficient and secure multi-party sum protocol [15]

This section presents the efficient and secure multi-party sum protocol in our previous work [15] (see in Fig. 1) that is employed as a main component of the proposed PPNBC solution.

Note that p and q are two large primes such that $q|(p-1)$, and g is an element in $\mathbb{Z}_p$ satisfying $g \neq 1$ and $g^q \bmod p = 1$. All computations in the protocol [15] are performed in $\mathbb{Z}_p$.

Input: $\boldsymbol{n}$ users $\{\boldsymbol{U_1}, \dots, \boldsymbol{U_n}\}$, each $\boldsymbol{U_i}$ holds a secret value $\boldsymbol{v_i} \in \{\boldsymbol{0}, \boldsymbol{1}\}$.
Output: the miner obtains $\boldsymbol{V} = \sum_{i=1}^{n} \boldsymbol{v_i}$, while the users do not reveal their private values with anyone.

Step 1: Each user $\boldsymbol{U_i}$ chooses two private keys $\boldsymbol{x_i}, \boldsymbol{y_i} \in [\boldsymbol{1}, \boldsymbol{q-1}]$, and computes the public keys $\boldsymbol{X_i} = \boldsymbol{g^{x_i}}$ & $\boldsymbol{Y_i} = \boldsymbol{g^{y_i}}$, then he/she shares $\boldsymbol{X_i}, \boldsymbol{Y_i}$ for the miner.
Step 2: The miner computes the shared public keys $\boldsymbol{X} = \prod_{i=1}^{n} \boldsymbol{X_i}$ & $\boldsymbol{Y} = \prod_{i=1}^{n} \boldsymbol{Y_i}$, then sends $\boldsymbol{X}, \boldsymbol{Y}$ to all users.
Step 3: Each user $\boldsymbol{U_i}$ encrypts his/her private value by computing $\boldsymbol{P_i} = \boldsymbol{g^{v_i}}.\boldsymbol{X^{y_i}}.\boldsymbol{Y^{q-x_i}}$, then sends $\boldsymbol{P_i}$ to the miner.
Step 4: The miner aggregates $\boldsymbol{K} = \prod_{i=1}^{n} \boldsymbol{P_i}$, and computes $\boldsymbol{V}$ that satisfies $\boldsymbol{g^V} = \boldsymbol{K}$.

Fig. 1. The efficient and secure protocol for computing sum values [15]

III. Privacy-preserving naïve bayes classification in fully distributed data setting

A. System initialization

The parameters (p, q, g) mentioned in Section II.0 are known by all participants.

As described in Figure 1, each private value is encrypted by employing two private keys. Moreover, Section II.A showed that the total number of frequency values computed by the miner is $T = t + t\sum_{j=1}^{m} l_j$. Hence, to jointly compute T frequency values mentioned in Section II.0, each data owner only chooses s private keys with $s = \left\lceil \frac{1}{2} + \sqrt{2T + \frac{1}{4}} \right\rceil$. In the other words, by employing T different tuples of private keys picked from the set of s private keys, each data owner can collaboratively compute T frequency values.

B. A privacy-preserving protocol for Naïve Bayes classification in the fully distributed data setting

Our proposed PPNBC solution includes four phases as described in 0.

Input: $\boldsymbol{n}$ data owners $\{\boldsymbol{U_1}, \dots, \boldsymbol{U_n}\}$, each $\boldsymbol{U_i}$ holds a private binary vector $\boldsymbol{u_i}$ of $(\boldsymbol{m+1})$ elements.
Output: the miner obtains $\boldsymbol{T}$ frequency values while the data owners do not disclose their private vectors with anyone.

Phase 1: Key preparation
- Each $\boldsymbol{U_i}$ chooses $\boldsymbol{s}$ private keys in $[\boldsymbol{1}, \boldsymbol{q-1}]$.
- Each $\boldsymbol{U_i}$ computes $\boldsymbol{s}$ corresponding public keys and sends them to the miner.

Phase 2: Shared public keys computation
- The miner computes the corresponding $\boldsymbol{s}$ shared public keys from $\boldsymbol{s}$ public keys received from each data owner.
- The miner sends $\boldsymbol{s}$ shared public keys to all data owners.

Phase 3: Data encryption
- For each of $\boldsymbol{T}$ frequency value, each $\boldsymbol{U_i}$ does:
  - ✧ Determines the corresponding private input $\boldsymbol{v_i}$
  - ✧ Picks a tuple of private keys (denoted as $\boldsymbol{x_i}, \boldsymbol{y_i}$) and the corresponding tuple of shared public keys (denoted as $\boldsymbol{X}, \boldsymbol{Y}$).
  - ✧ Computes the ciphertext $\boldsymbol{P_i} = \boldsymbol{g^{v_i}}.\boldsymbol{X^{y_i}}.\boldsymbol{Y^{q-x_i}}$.
- Each $\boldsymbol{U_i}$ sends all ciphertexts $\boldsymbol{P_i}$ to the miner.

Phase 4: Results extraction

Input: $\boldsymbol{n}$ data owners $\{\boldsymbol{U_1}, \ldots, \boldsymbol{U_n}\}$, each $\boldsymbol{U_i}$ holds a private binary vector $\boldsymbol{u_i}$ of $(\boldsymbol{m}+\boldsymbol{1})$ elements.
Output: the miner obtains $\boldsymbol{T}$ frequency values while the data owners do not disclose their private vectors with anyone.

- For each of $\boldsymbol{T}$ frequency value, the miner does:
  - ✧ Aggregates $\boldsymbol{K} = \prod_{i=1}^{n} \boldsymbol{P_i}$.
  - ✧ Outputs a frequency value $\boldsymbol{V}$ that satisfies $\boldsymbol{g^V} = \boldsymbol{K}$.
- The miner computes the necessary probabilities for Naïve Bayes classification from the frequency values.

Figure 2. The privacy-preserving protocol for Naïve Bayes classification in the fully distributed setting

C. Privacy analysis

It can be seen in 0 that our privacy-preserving protocol for Naïve Bayes classification in the fully distributed setting is composed of T secure multi-party sum protocols. Furthermore, the secure multi-party sum protocol's privacy was recognized in [15]. Thus, based on the security definition of a secure cryptographic protocol following the semi-honest model and the composition theorem mentioned in the book [16], the proposed privacy-preserving Naïve Bayes classification solution is semantically secure.

D. Accuracy analysis

Because the secure multi-party sum protocol's correctness was proved in [15], the protocol presented in 0 accurately computes the frequency values. Briefly, the Naïve Bayes classification model's accuracy is guaranteed in our proposal.

E. Efficiency evaluation

To show the efficiency of the proposed solution, this section compares the running time among three typical privacy-preserving Naive Bayes classifiers, i.e., the solution of Yang et al. in [6], the solution based on the secure e-voting protocol [17] of Hao et al. (named Yang's solution and Hao's-based solution, respectively), and ours. These PPNBC solutions are chosen for our comparisons, because they have the capability to ensure the accuracy of classification model, as well as the same high level of privacy.

Particularly, we consider the total running time of the miner and each data owner in the compared solutions when tested on the pre-processed German credit dataset [18] at UCI Machine Learning repository.

Our experiments are implemented in Python language and run on the virtual machine with Ubuntu operating system, 2 cores of the Intel core $i5-8250U$ @$1.6GHz$ CPU, 4 threads, and $4GB$ memory.

The experimental results are presented in TABLE I. It can be seen that Yang's, Hao's-based solutions, and ours are approximately the same total running time for each data owner (i.e., 1.308 seconds, 1.296 seconds, 1.312 seconds, respectively). For the miner, the total running time for him/her in our solution is only 2.666 seconds, while that in Hao's-based solution is 8.244 seconds. Especially, Yang's solution requires the miner up to 159.217 seconds to perform his/her tasks.

Additionally, the number of private keys used in Yang's and our solution is much less than the one in Hao's-based solution.

TABLE I. The running time (in seconds) comparison among our proposal and the typical PPNBC solutions

Solutions	Each data owner	The miner
Yang's solution	1.308	159.217
Hao's-based solution	1.296	8.244
Our solution	1.312	2.666

In summary, the above experimental results show that the proposed PPNBC solution is more efficient than the typical others. Thus, our solution is suitable for practical applications.

IV. Conclusion

In this work, we proposed an efficient method based on a secure multi-party sum protocol for privacy-preserving Naïve Bayes classification in the fully distributed data setting. Our proposed PPNBC solution not only protects each data owner's privacy but also guarantees the classification model's accuracy. The

experimental results have demonstrated the efficacy of our proposal. In the future, we will investigate the issue of privacy preservation for other machine learning techniques in various data models.

REFERENCES

[1] Y. Lindell and B. Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining," J. Priv. Confidentiality, vol. 1, no. 1, pp. 59–98, 2009, doi: https://doi.org/10.29012/jpc.v1i1.566.

[2] M. Kantarcıo˘glu, J. Vaidya, and C. Clifton, "Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data," presented at the IEEE ICDM workshop on privacy preserving data mining, 1-7, 2003. [Online]. Available: http://www.cis.syr.edu/~wedu/ppdm2003/papers/1.pdf

[3] J. Vaidya, M. Kantarcioglu, and C. Clifton, "Privacy-preserving Naïve Bayes classification," VLDB J., vol. 17, pp. 879–898, 2008, doi: https://doi.org/10.1007/s00778-006-0041-y.

[4] B. Schneier, Applied Cryptography, 2nd ed. John Wiley & Sons, 1996.

[5] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explor. Newsl., vol. 4, no. 2, pp. 28–34, 2002, doi: https://doi.org/10.1145/772862.772867.

[6] Z. Yang, S. Zhong, and R. N. Wright, "Privacy-Preserving Classiflcation of Customer Data without Loss of Accuracy," in Proceedings of the 2005 SIAM International Conference on Data Mining, 2005, pp. 92–102. doi: https://doi.org/10.1137/1.9781611972757.9.

[7] X. Yi and Y. Zhang, "Privacy-preserving Naive Bayes classification on distributed data via semi-trusted mixers," Inf. Syst., vol. 34, pp. 371–380, 2009, doi: https://doi.org/10.1016/j.is.2008.11.001.

[8] M. E. Skarkala, M. Maragoudakis, S. Gritzalis, and L. Mitrou, "PPDM-TAN: A Privacy-Preserving Multi-Party Classifier," Computation, vol. 9, no. 6, pp. 1–25, 2021, doi: https://doi.org/10.3390/computation9010006.

[9] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes," in International Conference on the Theory and Applications of Cryptographic Techniques, 1999, pp. 223–238. doi: https://doi.org/10.1007/3-540-48910-X_16.

[10] C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proceedings of the forty-first annual ACM symposium on Theory of computing, 2009, pp. 169–178. doi: https://doi.org/10.1145/1536414.1536440.

[11] P. Li, J. Li, Z. Huang, C.-Z. Gao, W.-B. Chen, and K. Chen, "Privacy-preserving outsourced classification in cloud computing," Clust. Comput., vol. 21, pp. 277–286, 2018, doi: https://doi.org/10.1007/s10586-017-0849-9.

[12] M. Huai, L. Huang, W. Yang, L. Li, and M. Qi, "Privacy-preserving Naive Bayes classification," in International conference on knowledge science, engineering and management, 2015, pp. 627–638. doi: https://doi.org/10.1007/978-3-319-25159-2_57.

[13] T. Li, J. Li, Z. Liu, P. Li, and C. Jia, "Differentially private Naive Bayes learning over multiple data sources," Inf. Sci., vol. 444, pp. 89–104, 2018, doi: https://doi.org/10.1016/j.ins.2018.02.056.

[14] P. Li, T. Li, H. Ye, J. Li, X. Chen, and Y. Xiang, "Privacy-preserving machine learning with multiple data providers," Future Gener. Comput. Syst., vol. 87, pp. 341–350, 2018, doi: https://doi.org/10.1016/j.future.2018.04.076.

[15] V. Duy Hien, L. The Dung, and H. Tu Bao, "An efficient approach for secure multi-party computation without authenticated channel," Inf. Sci., vol. 527, pp. 356–368, 2020, doi: https://www.doi.org/10.1016/j.ins.2019.07.031.

[16] O. Goldreich, "Basic Applications," in Foundations of Cryptography, vol. II, Cambridge University Press, 2004.

[17] F. Hao, P. Y. A. Ryan, and P. Zielin´ski, "Anonymous voting by two-round public discussion," IET Inf. Secur., vol. 4, no. 2, pp. 62–67, 2010, doi: https://doi.org/10.1049/iet-ifs.2008.0127.

[18] H. Hofmann, "Statlog (German Credit Data) Data Set," 1994. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

ABOUT THE AUTHORS

Vu Duy Hien
Workplace: Banking Academy of Vietnam.
Email: hienvd@hvnh.edu.vn.
Education: Master of Sciences.

Recent research direction: privacy-preserving data mining and machine learning, banking technology, public-key cryptography.

Luong The Dung
Workplace: Academy of Cryptography Techniques.
Email: thedungluong1@gmail.com
Education: Received Bachelor's degree in 2001, and PhD in 2011 in Mathematical foundation for computers and computing systems from Military Technology Academy

Recent research direction: data privacy, cryptographic, data mining, machine learning for information security

Hoang Duc Tho
Workplace: Academy of Cryptography Techniques.
Email: thohd80@gmail.com
Education: Received Bachelor's and Master's degrees in Information Technology from University of Engineering and Technology in 2002 and 2007; PhD in Information Security at FSO Academy - Russian Federation in 2014.

Recent research direction: cryptography information security