ISSN:0975-9646

A Survey on Classification over Semantically Secure Encrypted Data

Chaitali R. Shewale[1]

[1] *Computer Department, SP Pune University ,*
Pune Maharashtra, India

Abstract— **with the rapid development of web services and their popularity, web users are increasing day by day. As a result, there is a large and heterogeneous data. This data needs to be mine for various real time and other type of applications like banking, medicine, scientific research and among government agencies. Data Mining is a wider area now days due to the necessity of knowledge discovery from a different perspective on a very large scale. One of the commonly used tasks in data mining applications is the Classification. Many theoretical and practical solutions to the classification problem have been proposed in the past decades. To overcome the privacy issues, these solutions have used different security models. The current needs of IT makes the Cloud Computing comes into existence. Users can outsource their data in encrypted form and the data mining tasks to the cloud. Existing privacy preserving classification techniques are not applicable to the encrypted data on cloud. Therefore, this paper aims to solve the classification problem on encrypted data. A k-NN classifier over encrypted data in the cloud is proposed here for security purpose. The system proposes a protocol to provide the confidentiality of the data on the cloud protects privacy of user input query and hides the data access patterns on the cloud. A reliable k-NN classifier is the first over the encrypted data under the semi-honest model.**

Keywords— **Security, k-NN Classifier, Outsourced Databases, Encryption.**

I. INTRODUCTION

Recently, the cloud computing paradigm [1] is transfiguring the organizations' way of operating (i. e. store, access and process) their data. As an emerging computing paradigm, cloud computing provides the advantageous cloud potentials to the organizations in terms of its cost-efficiency, flexibility, and offload of administrative overhead. Most often, organizations represents their computational operations in addition to their data on the cloud. Cloud provides many advantages to the organizations, but the privacy and security issue in the cloud prevents companies to get those advantages. The highly sensitive data needs to be encrypted before outsourcing to the cloud. When data is encrypted, performing any data mining tasks becomes very difficult without ever decrypting the data.

There are other privacy concerns also, Consider an insurance company outsourced its customers database which is in encrypted form and relevant data mining tasks to a cloud. An agent from the company wants to estimate the risk level of a potential new customer; the agent uses a classification method to estimate the risk level of the customer. To do this, basically the agent needs to create a data record r for the customer that contains some personal information of the customer, like credit score, age, marital status, etc. Then this record is sent to the cloud, and the cloud has to compute the class label for r. Since, r contains sensitive information, to secure the user privacy; r should be encrypted before sending it to the cloud. The above scenario shows that data mining over encrypted data (denoted by DMED) on a cloud. As the record is a part of the data mining process it must need to protect. Moreover, cloud can also learn useful and sensitive information about the actual data items by observing the data access patterns even if the data is provided with the encryption [2], [3]. Therefore, the privacy/security requirements of the DMED issue on a cloud are: (1) confidentiality of the encrypted data, (2) confidentiality of a user's query record, and (3) hiding data access patterns.

Traditional methods on Privacy-Preserving Data Mining cannot solve the DMED problem. Because many intermediate computations while mining, are based on non encrypted data. As a result, this paper proposes novel methods to effectively solve the DMED problem. It is assumed that the encrypted data are outsourced to a cloud. To be more specific, we focus on the classification problem as it is one of the most common data mining tasks. This paper concentrates on executing the k-nearest neighbour classification process over encrypted data in the cloud computing environment.

The rest of the paper is arranged as follows: Section II presents the literature survey over the related work. In section III, proposed system is presented. Finally, the section IV concludes the review paper.

II. LITERATURE SURVEY

In this section, we have studied previous research papers related to the privacy preserving data mining (PPDM) and query processing over encrypted data. The brief review of existing related work is as follows:

C. Gentry [4] presents a fully homomorphic cryptosystems to solve the DMED problem. It allows a third-party (that hosts the encrypted data) to execute random functions over encrypted data without ever decrypting them. The problem with this system is, such techniques are very costly and they are not yet practically explored. For example, C. Gentry and S. Halevi [5] shows that even for inadequate security parameters one "Loading" operation on the homomorphic system takes a time on a high performance machine. It takes at least 30 seconds to complete the task.

A. Shamir [6] proposed a secret sharing scheme in secure multiparty computation (SMC), to develop a PPkNN protocol. SMC based approach assumes data are divided and it is not encrypted at each participating party, intermediate computations are performed on non-encrypted data. Opposite to this our proposed work is different from the secret sharing based solution. Let us see how it is. Secret sharing based methods involves minimum three parties in the system whereas our system will work using only two parties. So, in our system there is a less chance of the deception of the data security.

For example, D. Bogdanov, S. Laur, and J. Willemson [7] proposed the constructions based on Sharemind, a well-known SMC framework which is based on the secret sharing scheme, it assumes that the number of participating parties is three. Thus, our work is orthogonal to Sharemind and other secret sharing based schemes

A Privacy-Preserving Data Mining (PPDM)

R. Agrawal and R. Srikant [8], Y. Lindell and B. Pinkas [9] introduce the notion of privacy preserving under data mining applications. The traditional PPDM techniques can be divided into two categories: (i) data perturbation and (ii) data distribution. Agrawal and Srikant [8] proposed the first data perturbation technique to build a decision-tree classifier, later many other methods were proposed. However, data perturbation techniques cannot be applicable for semantically protected encrypted data. Due to the addition of statistical noises to the data, perturbation techniques do not produce accurate data mining results. On the other hand, Lindell and Pinkas [9] proposed the first decision tree classifier under the two-party setting. The user does not have real keywords to describe the queries assuming the data were distributed between them. Since then much work has been published using SMC techniques. It is studied that the PPkNN problem cannot be solved using the data distribution techniques as the data in our case is encrypted and not distributed in plaintext among multiple parties. Hence, we also do not consider secure k-NN methods in which the data are distributed between two parties.

B Query Processing over Encrypted Data

There are many techniques related to query processing over encrypted data have been proposed, in [17]–[19]. It is studied that PPkNN is a more complex problem than the running of simple kNN queries over encrypted data [20], [21]. For one, the transitional k-nearest neighbours in the classification process should not be revealing to the cloud or any other users unlike the recent method in [21] which reveals the k-nearest neighbours to the user. Secondly, even if we know the k-nearest neighbours, it is still very hard to find the majority class label between these neighbours since they are encrypted at the first phase to prevent the cloud from learning sensitive information. Third, the existing work does not overcome the access pattern issue which is a key privacy need from the user's perspective.

Y. Elmehdwi, B. K. Samanthula [22], proposed a novel secure k-nearest neighbour query protocol over encrypted data. This protocol protects data confidentiality, user's query privacy, and hides data access patterns. However, as mentioned above, PPkNN is a more

Composite problem and it cannot be solved directly using the existing secure k-nearest neighbour techniques over encrypted data. Therefore, this paper provides a new solution to the PPkNN classifier problem over encrypted data.

More specifically, this paper is different from the above existing work [22] in the following three aspects. First, this paper, introduces new security primitives, namely secure minimum (SMIN), secure minimum out of n numbers (SMINn), secure frequency (SF), and found new solutions for them. Second, the work in [22] did not provide any formal security analysis of the underlying sub-protocols. On the contrary, this paper provides formal security proofs of

The underlying sub-protocols and the PPkNN protocol under the semi-honest model. Third, the preliminary work in [22] addresses only secure kNN query which is similar to Stage 1 of PPkNN. However, Stage 2 in PPkNN is entirely new.

III. PROPOSED SYSTEM

Existing privacy preserving classification techniques are not applicable to the encrypted data on cloud. Therefore, this paper aims to solve the classification problem on encrypted data. A k-NN classifier over encrypted data in the cloud, is proposed here for security purpose. The system proposes a protocol to provide the confidentiality to the encrypted data on the cloud, protects privacy of user input query and hides the data access patterns on the cloud. A reliable k-NN classifier is the first over the encrypted data under the semi-honest model. PPkNN is a more composite problem and it cannot be solved directly using the existing secure k-nearest neighbour techniques over encrypted data. Therefore, this paper proposed a new solution to the PPkNN classifier problem over encrypted data. For this, proposed system uses a set of generic sub-protocols that will be used in constructing the proposed k-NN classifier. The security primitives used by the protocols are: i) secure minimum (SMIN), ii) secure minimum out of n numbers (SMINn) and iii) secure frequency (SF). Proposed system provides solution to each of these primitive.

Advantages of proposed protocol/system:

1. Protects the confidentiality of data
2. Privacy of user's input query
3. Hide the data access patterns.

IV. CONCLUSIONS

This paper reviews various existing methods used for the privacy preserving data mining (PPDM) and query processing over encrypted data. To protect user privacy, various privacy-preserving classification techniques have been proposed over the past decade. The existing techniques are not applicable to outsourced database environments where the data resides in encrypted form on a third-party server. This paper proposed a new privacy-preserving k- NN classification protocol over encrypted data in the cloud. This protocol protects the confidentiality of the data, user's input query, and hides the data access patterns.

Future research can be focus on more efficient solutions to the SMINn problem, because the performance of the PPkNN protocol depends on the efficiency of the SMINn. Also, this paper can be extended towards the other classification algorithms.

ACKNOWLEDGMENT

I am glad to express my sentiments of gratitude to all who rendered their valuable guidance to me. I would also love to thank my guide Prof. Helly Patel for her guidance. I would like to express my appreciation and thanks to the Principal of our college. I am also thankful to the Head of Department. I thank to the anonymous reviewers for their valuable comments.

REFERENCES

[1] P. Mell and T. Grance, "The nist definition of cloud computing (draft)," NIST special publication, vol. 800, p. 145, 2011.

[2] S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "Managing and accessing data in the cloud: Privacy risks and approaches," in CRiSIS, pp. 1 –9, 2012.

[3] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: practical access pattern privacy and correctness on untrusted storage," in ACM CCS, pp. 139–148, 2008.

[4] C. Gentry, "Fully homomorphic encryption using ideal lattices," in ACM STOC, pp. 169–178, 2009

[5] C. Gentry and S. Halevi, "Implementing gentry's fullyhomomorphic encryption scheme," in EUROCRYPT, pp. 129–148, Springer, 2011.

[6] A. Shamir, "How to share a secret," Commun. ACM, vol. 22, pp. 612–613, Nov. 1979.

[7] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations," in ESORICS, pp. 192–206, Springer, 2008.

[8] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in ACM Sigmod Record, vol. 29, pp. 439–450, ACM, 2000.

[9] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Advances in Cryptology (CRYPTO), pp. 36–54, Springer, 2000.

[10] P. Zhang, Y. Tong, S. Tang, and D. Yang, "Privacy preserving naive bayes classification," ADMA, pp. 744–752, 2005.

[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Information Systems, vol. 29, no. 4, pp. 343–364, 2004.

[12] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in IEEE ICDE, pp. 217–228, 2005.

[13] H. Hu, J. Xu, C. Ren, and B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism," in IEEE ICDE, pp. 601–612, 2011.

[14] M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier," in PKDD, pp. 279–290, 2004.

[15] L. Xiong, S. Chitti, and L. Liu, "K nearest neighbor classification across multiple private databases," in CIKM, pp. 840–841, ACM, 2006.

[16] Y. Qi and M. J. Atallah, "Efficient privacy-preserving k-nearest neighbor search," in *IEEE ICDCS*, pp. 311–319, 2008.

[17] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in ACM SIGMOD, pp. 563–574, 2004.

[18] H. Hacig¨um¨ us¸, B. Iyer, C. Li, and S. Mehrotra, "Executing sql over encrypted data in the database-service-provider model," in *ACM SIGMOD*, pp. 216–227, 2002.

[19] B. Hore, S. Mehrotra, M. Canim, and M. Kantarcioglu, "Secure multidimensional range queries over outsourced data," The VLDB Journal, vol. 21, no. 3, pp. 333–358, 2012.

[20] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure knn computation on encrypted databases," in ACM SIGMOD, pp. 139–152, 2009.

[21] X. Xiao, F. Li, and B. Yao, "Secure nearest neighbor revisited," in IEEE ICDE, pp. 733–744, 2013.

[22] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, "Secure k- nearest neighbor query over encrypted data in outsourced environments," in IEEE ICDE, pp. 664–675, 2014.