ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A SURVEY ON PRIVACY PRESERVING TECHNIQUE USING K- NEAREST NEIGHBOR CLASSIFICATION

Rashmi Sheshrao Kodane*, Prof. Karuna Bagde
Student of Master of Engineering in (CS & IT) HVPM's college of Engineering and Technology Amravati, India
Associate Professor in Department of (CSE) HVPM's College of Engineering and Technology Amravati, India

DOI: 10.5281/zenodo.47037

ABSTRACT
Data mining is a powerful new technique to discover knowledge within the large amount of the data. Also data mining is the process of discovering meaningful new relationship, patterns and trends by passing large amounts of data stored in corpus, using pattern recognition technologies as well as statistical and mathematical techniques. 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 novel privacy-preserving k-NN classification protocol over encrypted data in the cloud. Our protocol protects the confidentiality of the data, user's input query, and hides the data access patterns. We also evaluated the performance of our protocol under different parameter settings.

KEYWORDS: Security, k-NN classifier, outsourced databases, encryption, privacy preserving.

INTRODUCTION
Data Mining has wide applications in many areas such as banking, medicine, scientific research and among government agencies. Classification is one of the commonly used tasks in data mining applications. For the past decade, due to the rise of various privacy issues, many theoretical and practical solutions to the classification problem have been proposed under different security models. However, with the recent popularity of cloud computing, users now have the opportunity to outsource their data, in encrypted form, as well as the data mining tasks to the cloud. Since the data on the cloud is in encrypted form, existing privacy-preserving classification techniques are not applicable. In this paper, we focus on solving the classification problem over encrypted data. In particular, we propose a secure k-NN classifier over encrypted data in the cloud. The proposed protocol protects the confidentiality of data, privacy of user's input query, and hides the data access patterns. Despite tremendous advantages that the cloud offers, privacy and security issues in the cloud are preventing companies to utilize those advantages. When data are highly sensitive, the data need to be encrypted before outsourcing to the cloud. However, when data are encrypted, irrespective of the underlying encryption scheme, performing any data mining tasks becomes very challenging without ever decrypting the data.

Data mining over encrypted data (denoted by DMED) on a cloud needs to protect a user's record when the record is a part of a data mining process. Moreover, cloud can also derive useful and sensitive information about the actual data items by observing the data access patterns even if the data are encrypted. Therefore, the privacy/security requirements of the DMED problem on a cloud are threefold: (1) confidentiality of the encrypted data, (2) confidentiality of a user's query record, and (3) hiding data access patterns. Existing work on privacy-preserving data mining (PPDM) (either perturbation or secure multi-party computation (SMC) based approach) cannot solve the DMED problem. Perturbed data do not possess semantic security, so data perturbation techniques cannot be used to encrypt highly sensitive data. Also the perturbed data do not produce accurate data mining results. Secure multi-party computation based approach assumes data are distributed and not encrypted at each participating party. In addition, many intermediate computations are performed based on non-encrypted data.

As a result, in this, we proposed novel methods to effectively solve the DMED problem assuming that the encrypted data are outsourced to a cloud. Specifically, we focus on the classification problem since it is one of the most common data mining tasks. Because each classification technique has their own advantage, to be concrete, this paper concentrates on executing the k-nearest neighbor classification method over encrypted data in the cloud computing environment.

LITERATURE SURVEY

According to P. Mell and T. Grance, The cloud computing paradigm is revolutionizing the organizations' way of operating their data particularly in the way they store, access and process data [1] . Moreover, cloud can also derive useful and sensitive information about the actual data items by observing the data access patterns even if the data are encrypted access patterns [2] [3]. Lindell and Pinkas were the first to introduce the notion of privacy - preserving under data mining applications. The existing PPDM techniques can broadly be classified into two categories: (i) data perturbation and (ii) data distribution. P. Williams, R. Sion, and B. Carbunar [3] author provide novel and effective practical scheme with efficient access pattern privacy for remote data storage with correctness. Major aim of invention of this protocols to yield practical computational complexity (to O(log n log log n)) and storage overheads (to O(n)). Storage client issue encrypted reads, writes without revealing information or access patterns by using proposed mechanism. Proposed scheme is faster than existing system which can execute several queries per second and also offering privacy as well as correctness. C. Gentry [4], author proposed a fully homomorphic encryption scheme based on concept of lattice. Proposed work consists of three steps; initial step provide a general result of encryption scheme which permits for evaluation of arbitrary circuit, then a public key encryption scheme based on concept of lattices and final step is to reduce the depth of decryption circuit and used to produce a bootstrappable encryption scheme. Proposed scheme solve the DMED problem since it allows a third-party to execute arbitrary functions over encrypted data without ever decrypting them. This technique is very expensive and their usage in practical applications has yet to be explored. Agrawal and Srikant[5], Lindell and Pinkas [6], introduced the idea of privacy-preserving under data mining applications. Various techniques related to query processing over encrypted data have been proposed, e.g., [7], [8], [9]. However, we observe that PPkNN is a more complex problem than the execution of simple kNN queries over encrypted data [10], [11]. For one, the intermediate k-nearest neighbors in the classification process, should not be disclosed to the cloud or any users. We emphasize that the recent method in [11] reveals the k-nearest neighbors to the user. R. Dey, C. Tang, K. Ross, and N. Saxena, paper mainly focus on solving the problem of encrypted data classification. Paper proposed a novel PPkNN protocol, a secure k-NN classifier over semantically secure encrypted data. Author proposed a secure k-NN classifier for encrypted data in the cloud. Commonly used scheme in data mining is classification which is used in health-care and business.. The proposed KNN (k Nearest Neighbor) protocol provides protection for the users input query, confidentiality of the data and data access patterns. Efficiency of proposed method gives better result.

In proposed protocol once the encrypted data are handover to the cloud, Alice does not participate in any computations. Thus, no information is revealed to Alice which indirectly achieves privacy. Y. Elmehdwi, B. K. Samanthula [12], 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 [12] 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 [12] 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 [12] addresses only secure kNN query which is similar to Stage 1 of PPkNN. However, Stage 2 in PPkNN is entirely new.

In our most recent work we proposed a novel secure k-nearest neighbor query protocol over encrypted data that protects data confidentiality, user's query privacy, and hides data access patterns. However, as mentioned above, PPkNN is a more complex problem and it cannot be solved directly using the existing secure k-nearest neighbor

techniques over encrypted data. Therefore, in this seminar, we extend our previous work and provide a new solution to the PPkNN classifier problem over encrypted data. On the other hand, this paper provides formal security proofs of the underlying sub-protocols as well as the PPkNN protocol under the semi-honest model.

COMPARISON OF VARIOUS ENCRYPTION SCHEME WITH K-NEAREST NEIGHBOR CLASSIFICATION

Table 1. Comparison of various Encryption scheme with k-nearest neighbour classification

Methods	Advantages	Disadvantages
k-Nearest Neighbor Classification with Traditional Encryption Scheme	1. Data are secured 2. Classification error is very less due to decryption	1. More damage if compromised 2. Sharing the private key 3. If the data size is large then processing speed will become slow 4 Encryption and Decryption over head is very high
Privacy Preserving k-Nearest Neighbor Classification (PPkNN)with homomorphic encryption	1. K-Nearest Neighbor classification to be carried out the encrypted data set 2. Only encryption method is used, is to reduce the overhead 3. Not sharing secret key 4. Sensitive data are more secured 5. Computing class labels rapidly	1. Encryption is done by using only partial homomorphic scheme

PROPOSED WORK

In this seminar, we propose a novel PPkNN protocol, a secure k-NN classifier over semantically secure encrypted data. So first we study the existing technique such as centralized and distributed data and try to achieve this problem by using k-NN classification using paillier cryptosystem using semi-honest model. First, in this we introduced new security primitives, namely secure minimum (SMIN), secure minimum out of n numbers (SMINn), secure frequency (SF), and proposed new solutions for them. Second, the work did not provide any formal security analysis of the underlying sub-protocols. We provides formal security proofs of the underlying sub-protocols as well as the PPkNN protocol under the semi-honest model. We also compare propose a privacy-preserving technique using kernel density estimation using a Gaussian kernel, a classication algorithm from the same family as k-NN. We additionally investigate solutions for other threat models, often through extensions on prior single data owner systems. We can also propose k-NN classification using other data mining techniques like: K-means algorithm and Bayesian algorithm. So we discuss the following technique:

[1] Privacy-Preserving Primitives
[2] Security Analysis of Privacy-Preserving Primitives under the Semi-honest model
[3] The Proposed PPkNN Protocol
[4] Security Analysis of PPkNN under the Semi-Honest Model

PROPOSED TECHNIQUE:

Cryptography based PPDM (Privacy Preserving Data Mining)

This technique includes secure multiparty computation where a computation is secure if at the completion of the computation, no one can know anything except its own input and the results. Cryptographybased algorithms are considered for protective privacy in a distributed situation by using encryption techniques. Transformed data are exact and protected. Better privacy compare to randomized approach. It offers a well-defined model for privacy that includes methods for proving and quantifying it. Second a vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain.

Advantage ofCryptography based PPDM

1)Because the process is transparent, it is easy to implement and debug.
2)In situations where an explanation of the output of the classifier is useful Cryptography based PPDMcan be very effective if an analysis of the neighbors is useful as explanation.
3)There are some noise reduction techniques that work only for Cryptography based PPDM that can be effective in improving the accuracy of the classifier.

CONCLUSION

To protect user privacy, various privacy- preserving classification techniques have been proposed over the past decade. The existing techniques are not applicable to out sourced database environment s where the data resides in encrypted form on a third- party server. This paper proposed a novel privacy- preserving k-NN classification protocol over encrypted data in the cloud. Our protocol protects the confidentiality of the data, user's input query, and hides the data access pat terns. We also evaluated the performance of our protocol under different parameter set tings. Since improving the efficiency of SMIN n is an import ant first step f or improving the -performance of our PPkNN protocol, we plan to investigate alternative and more efficient solute- ions t o t he SMIN n problem in our future work. Also, we will investigate and extend our research to other classification algorithms.

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 Proc. 7th Int. Conf. Risk Security Internet Syst., 2012, pp. 1–9.
[3] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: Practical access pattern privacy and correctness on untrusted storage," in Proc. 15th ACM Conf. Comput. Commun. Security, 2008, pp. 139–148.
[4] C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proc. 41st Annu. ACM Sympos. Theory Comput., 2009, pp. 169–178.
[5] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM Sigmod Rec., vol. 29, pp. 439–450, 2000.
[6] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Proc. 20th Annu. Int. Cryptol. Conf. Adv. Cryptol., 2000, pp. 36–54.
[7] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2004, pp. 563–574.
[8] H. Hacig€um€u ,s, B. Iyer, C. Li, and S. Mehrotra, "Executing sql over encrypted data in the database-service-provider model," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2002, pp. 216–227.
[9] B. Hore, S. Mehrotra, M. Canim, and M. Kantarcioglu, "Secure multidimensional range queries over outsourced data," VLDB J., vol. 21, no. 3, pp. 333–358, 2012.
[10] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure kNN computation on encrypted databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 139–152.
[11] X. Xiao, F. Li, and B. Yao, "Secure nearest neighbor revisited," in Proc. IEEE Int. Conf. Data Eng., 2013, pp. 733–744.
[12] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted relational data,"
[13] eprint arXiv:1403.5001, 2014.