PERFORMANCE IMPROVEMENT OF CLOUD DATA USING KEYWORD SEARCH

[1]Khushboo Patel [2]Shriya Chole [3]Abhay Kumar [4]Rohini Padole [5]Anup Bhange
[1234] U. G. Student, Dept. of Computer Technology. KDKCE Nagpur, Maharashtra, India
5Assistant Professor Dept OF Computer Tech KDKDCE, Nagpur

Abstract — 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 hybrid k-NN classifier over encrypted data in the cloud. The propose hybrid k-NN protocol protects the confidentiality of the data, user's input query, and data access patterns. To the best of our knowledge, our work is the first to develop a secure k-NN classifier over encrypted data under the semi-honest model. Also, we empirically analyze the efficiency of our solution through various experiments

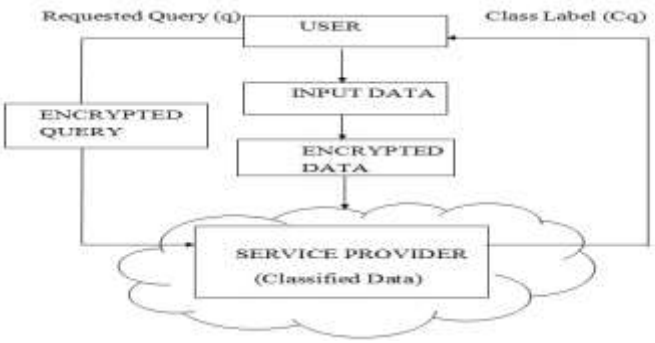
Keywords —k-NN,classifier,cloud,encryption ,mining,privacy preservation.

I. INTRODUCTION

Today's digital infrastructure supports innovative ways of storing, processing, and disseminating data. In fact, we can store our data in remote servers, access reliable and efficient services provided by third parties, and use computing power available at multiple locations across the network. Furthermore, the growing adoption of portable devices (e.g., PDAs, mobile phones) together with the diffusion of wireless connections in home and work environments have led to a more distributed computing scenario. These advantages come at a price of higher privacy risks and vulnerabilities as a huge amount of (private) information is being circulated and stored, often not under the direct control of its owner. Be that as it may, when information are encoded, independent of the fundamental encryption plan, performing any information mining assignments turns out to be extremely difficult without ever unscrambling the information. There are other sample security concerns, shown by the accompanying Data mining over encrypted data (denoted by DMED) [3] on a cloud also needs to protect a client's record when the record is a part of a data mining process. However cloud can also abstract useful and sensitive information about the outsource 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 of three types: (1) privacy of the encrypted data, (2) privacy of a user's query record, and (3) hiding data access patterns.

Privacy-Preserving Data Mining (PPDM)

Privacy Preserving Data Mining (PPDM) is defined as the process of extracting/deriving the knowledge about data without compromising the privacy of data [3, 41, 48]. In the past decade, many privacy-preserving classification techniques have been proposed in the literature in order to protect user privacy. Agrawal and Srikant [3], Lindell and Pinkas [40] introduced the notion of privacy-preserving under data mining applications. In particular to privacy preserving classification, the goal is to build a classifier in order to predict the class label of input data record based on the distributed training dataset without compromising the privacy of data. 1. Data Perturbation Methods: In these methods, values of individual data records are perturbed by adding random noise in a such way that the distribution of perturbed data look very different from that of actual data. After such a transformation, the perturbed data is sent to the miner to perform the desired data mining tasks. Agrawal and Srikant [3] proposed the first data perturbation technique to build a decision-tree classifier. Since then many other randomizationbased methods have been proposed in the literature such as [5]. However, as mentioned earlier in Section 1, data perturbation techniques cannot be applicable for semantically secure encrypted data. Also, they do not produce accurate data mining results due to the addition of statistical noises to the data. 2. Data Distribution Methods: These methods assume the dataset is partitioned either horizontally or vertically and distributed across different parties. The parties later can collaborate to securely mine the combined data and learn 4 the global data mining results. During this process, data owned by individual parties is not revealed to other parties. This approach was first introduced by Lindell and Pinkas [14] who proposed a decision tree classifier under two-party setting. Since then much work has been published using secure multiparty computation techniques [1, 15].



1. Data Perturbation Methods: In these methods, values of individual data records are perturbed by adding random noise in a such way that the distribution of perturbed data look very different from that of actual data. After such a transformation, the perturbed data is sent to the miner to perform the desired data mining tasks. Agrawal and Srikant [3] proposed the first data perturbation technique to build a decision-tree classifier. Since then many other randomizationbased methods have been proposed in the literature such as [5]. However, as mentioned earlier in Section 1, data perturbation techniques cannot be applicable for semantically secure encrypted data. Also, they do not produce accurate data mining results due to the addition of statistical noises to the data. 2. Data Distribution Methods: These methods assume the dataset is partitioned either horizontally or vertically and distributed across different parties. The parties later can collaborate to securely mine the combined data and learn 4 the global data mining results. During this process, data owned by individual parties is not revealed to other parties. This approach was first introduced by Lindell and Pinkas [14] who proposed a decision tree classifier under two-party setting. Since then much work has been published using secure multiparty computation techniques [1, 15]. Classification is one important task in many applications of data mining such as health-care and business. Recently, performing data mining in the cloud attracted significant attention. In cloud computing, data owner outsources his/her data to the cloud. However, from user's perspective, privacy becomes an important issue when sensitive data needs to be outsourced to the cloud. The direct way to guard the outsourced data is to apply encryption on the data before outsourcing. Unfortunately, since the hosted data on the cloud is in encrypted form in our problem domain, the existing privacy preserving classification techniques are not sufficient and applicable to PPkNN due to the following reasons. (i) In existing methods, the data are partitioned among at least two parties, whereas in our case encrypted data are hosted on the cloud. (ii) Since some amount of information is loss due to the addition of statistical noises in order to hide the sensitive attributes, the existing methods are not accurate. (iii) Leakage of data access patterns: the cloud can easily derive useful and sensitive information about users' data items by simply observing the database access patterns. For the same reasons, in this paper, we do not consider secure k-nearest neighbor techniques in which the data are distributed between two parties (e.g., [12]). 2.2 Query processing over encrypted data Using encryption as a way to achieve the data confidentiality may cause another issue at the cloud during the query evaluation. The question here is "how can the cloud perform computations over encrypted data while the data stored are in encrypted form?" Along this direction, various techniques related to query processing over encrypted data have been proposed, e.g., [12]. However, we observe that PPkNN is a more complex problem than the execution of simple KNN queries over encrypted data [13]. 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 [14] reveals the k-nearest neighbors to the user. Secondly, even if we know the k-nearest neighbors, it is still very difficult to find the majority class label among these neighbors since they are encrypted at the first place to prevent the cloud from learning sensitive information. Third, the existing work did not addressed the access pattern issue which is a crucial privacy requirement from the user's perspective. In our most recent work [12], 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 paper, we extend our previous work in and provide a new solution to the PPkNN classifier problem over encrypted data. More specifically, this paper is different from our preliminary work in the following four aspects. First, in this paper, 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 in did not provide any formal security analysis of the underlying sub-protocols. 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. Additionally, we demonstrate various techniques through which the proposed protocol can possibly be extended to a protocol that is secure under the malicious model. Third, our preliminary work in addresses only secure KNN query which is similar to Stage 1 of PPKNN. However, Stage 2 in PPKNN is entirely new. Finally, our empirical analyses in Section VI are based on a real dataset whereas the results in are based on a simulated dataset. In addition, new results are included in this paper. As mentioned earlier, one can implement the proposed protocols under secret sharing schemes. By doing so, we need to have at least three independent parties. In this work, we only concentrate on the two party situation ; thus, we adopted the Paillier cryptosystem. Two-party and multi-party (three or more parties) SMC protocols are complement to each other, and their applications mainly depend on the number of available participants. In practice, two mutually independent clouds are easier to find and are cheaper to operate. On the other hand, utilizing three cloud servers and secret sharing schemes to implement the proposed protocols may result more efficient running time. We believe both two-party and multi-party schemes are important. As a future work, we will consider secret sharing based PPKNN 5 implementations.**II.**

LITERATURE SURVEY

[1] 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

In this Data mining is the extraction of interesting patterns or knowledge from huge amount of data. In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. A number of methods and techniques have been developed for privacy preserving data mining. This paper provides a wide survey of different privacy preserving data mining algorithms and analyses the representative techniques for privacy preserving data mining, and points out their merits and demerits. Finally the present problems and directions for future research are discussed.

[2] 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.

This paper presents We present the reticent statistical zero-knowledge protocols to prove statements such as: A committed number is a prime .

- A committed (or revealed) number is the product of two safe primes, i.e., primes p and q such that $(p-1)=2$and $(q-1)=2$areprime.
- A given integer has large multiplicative order modulo a composite number that consists of two safe prime factors.

[3] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Eurocrypt , pp. 223–238, 1999

For the past decade, query processing on relational data has been studied extensively, and many theoretical and practical solutions to query processing have been proposed under various scenarios. With the recent popularity of cloud computing, users now have the opportunity to outsource their data as well as the data management tasks to the cloud. However, due to the rise of various privacy issues, sensitive data (e.g., medical records) need to be encrypted before outsourcing to the cloud. In addition, query processing tasks should be handled by the cloud; otherwise, there would be no point to outsource the data at the first place. To process queries over encrypted data without the cloud ever decrypting the data is a very challenging task. In this paper, we focus on solving the k-nearest neighbor (KNN) query problem over encrypted database outsourced to a cloud: a user issues an encrypted query record to the cloud, and the cloud returns the k closest records to the user. We first present a basic scheme and demonstrate that such a naive solution is not secure. To provide better security, we propose a secure kNNprotocol that protects the confidentiality of the data, user's input query, and data access patterns. Also, we empirically analyze the efficiency of our protocols through various experiments. These results indicate that our secure protocol is very efficient on the user end, and this lightweight scheme allows a user to use any mobile device to perform the KNN query

[4] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted re-lational data." eprint arXiv:1403.5001, 2014.

Ensuring proper privacy and protection of the information stored, communicated, processed, and disseminated in the cloud as well as of the users accessing such information is one of the grand challenges of our modern society. As a matter of fact, the advancements in the Information Technology and the diffusion of novel paradigms such as data outsourcing and cloud computing, while allowing users and companies to easily access high quality applications and services, introduce novel privacy risks of improper information disclosure and dissemination. In this paper, we will characterize different aspects of the privacy problem in emerging scenarios. We will illustrate risks, solutions, and open problems related to ensuring privacy of users accessing services or resources in the cloud, sensitive information stored at external parties, and accesses to such information.

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

Most of the cryptographic work in privacy-preserving distributed datamining deals with semi-honest adversaries, which are assumed to follow the prescribed protocol but try to infer private information using the messages they receive during the protocol. Although the semi-honest model is reasonable in some cases, it is unrealistic to assume that adversaries will always follow the protocols exactly. In particular, malicious adversaries could deviate arbitrarily from their prescribed protocols. Secure protocols that are developed against malicious adversaries require utilization of complex techniques. Clearly, protocols that can withstand malicious adversaries provide more security. However, there is an obvious trade-off: protocols that are secure against malicious adversaries are generally more expensive than those secure against semi-honest adversaries only. In this paper, our goal is to make an analysis of trade-offs between performance and security in privacy preserving distributed data mining algorithms in the two models. In order to make a realistic comparison, we enhance commonly used sub protocols that are secure in the semi-honest model with zero knowledge proofs to be secure in the malicious model. We compare the performance of these protocols in both models.

Name of author	Algorithm used	Disadvantages
1.S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "Managing and accessing data in the cloud: Privacy risks and approaches," in CRiSIS	the reticent statistical zero-knowledge protocols	Privacy preservation not considered.
2. P.Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Eurocrypt	Homomorphic encryption	Classification not consider
3.B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted re-lational data	"k-nearest neighbor classification	Searching operation is time consuming.
4.C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in EUROCRYPT , pp. 129– 148, Springer, 2011.	Fully homomorphic	Classification not consider.

[1]Khushboo Patel [2]Shriya Chole [3]Abhay Kumar [4]Rohini Padole [5]Anup Bhange
[1234] U. G. Student, Dept. of Computer Technology. KDKCE Nagpur, Maharashtra, India
5Assistant Professor Dept OF Computer Tech KDKDCE, Nagpur

IV. CONCLUSION

From the above literature survey it is clearly observed that Different type of privacy preserving classification has been introduced in past few years. This method is not applicable to outsourced databases. Here there is need to improve performance of data classification and searching over cloud encrypted data proposed system makes the system highly scalable and minimizes information leakage. Prevents overloads by ranking the files at the user side, reducing bandwidth and protects document frequency.we try to implement system is secure, scalable and accurate compared to the other ranked keyword search.

V. REFERENCES

[1] D.Nurmi, R.Wolski, C.Grzegorczyk, G.Obertelli,S.Soman, L.Youseff and D.Zagorodnov, "The eucalyptus open-source cloud-computing system," CCGRID 20009.9th IEEE/ACM International Symposium, 2009.

[2] S.S and A. Basu, "Performance of eucalyptus and open stack clouds on future grid,"International Journal of Computer Applications, vol. 80,no.13,pp.31-37, 2013.

[3] Z.Pantić and M. A.Babar, "Guidelines for Building a Private Cloud Infrastructure," IT University of Copenhagen, Denmark, Copenhagen, Denmark,2012.

[4] B. Beal, "Public vs. private cloud applications: twocriticaldifferences,"23May2012.[Online]. Available:http://searchcloudapplications.techtarget.com/feature/Public-vs-private-cloud-applicationsTwo-critical-differences.

[5] Tarik Moataz, Abdullatif Shikfa, "Boolean symmetric searchable encryption," ASIA CCS '13 Proc. of the 8th ACM SIGSAC symposium on Information computer and communications security, .pp. 265- 276, NY, USA , 2013.

[6] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, "SearchableSymmetric Encryption: Improved Definitions and Efficient Constructions,"Proc. ACM 13th Conf. Computer and Comm. Security (CCS), 2006

[7] D. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2000.

[8] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in Proc. of EUROCRYP'04, volume 3027 of LNCS. Springer, 2004.

[9] S.Adhikari,G.Bunce,W.Chan,A.Chandramouly,D.Kamhout, B.McGeough,J.JonSlusser,C.Spence and B. Sunderland, "Best practices for building and enterprise private cloud," Intel IT Centre,2011.

[10] B.Adler,"Designing Private and hybrid clouds: architectural best practices," RightScaleInc.,2012.

[11] "Planning Guide: Virtualisation and cloud computing," Intel IT Centre,2013

[12] Y.Wadia,"TheEucalyptusOpenSourcePrivateCloud. over Untrusted Data Cloud through Privacy Homomorphism,"Proc. IEEE 27th Int'l Conf. Data Eng. (ICDE), 2011.

[13] G.VonLaszewski,J.Diaz, F.WangandG.Fox, "Comparison of multiple cloud frameworks," IEEE on Cloud computing(CLOUD),vol.734,no.741,pp.24-29,2012,5th International Conference

[14] F. Bao, R. Deng, X. Ding, and Y. Yang, "Private query on encrypted data in multi-user settings," in Proc. of ISPEC 2008.

[15] A. Swaminathan, Y. Mao, G.-M. Su, H. Gou, A.L. Varna, S. He, M. 5u, and D.W. Oard, "Confidentiality-Preserving Rank-Ordered Search," Proc. Workshop Storage Security and Survivability, 2007.

[16] Cong Wang, Ning Cao, Jin Li, Kui Ren, Wenjing Lou, "Secure ranked keyword search over encrypted cloud data," IEEE 2010 30th International Conference 2010.

[17] S. Zerr, D. Olmedilla, W. Nejdl, and W. Siberski, "Zerber+r: Top-k Retrieval from a Confidential Index," Proc. 12th Int'l Conf.Extending Database Technology: Advances in Database Technology (EDBT), 2009.

[18] P. Golle, J. Staddon, and B. Waters, "Secure Conjunctive Keyword Search over Encrypted Data," over Untrusted Data Cloud through Privacy Homomorphism,"Proc. IEEE 27th Int'l Conf. Data Eng. (ICDE), 2011.

[19] L. Ballard, S. Kamara, and F. Monrose, "Achieving Efficient Conjunctive Keyword Searches over Encrypted Data.