IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

K-Nearest Neighbor Classification on Data Confidentiality and, Privacy of User's Input Queries

Dr. V. Goutham[1], P. Ashwini Reddy[2], K. Sunitha[3]

Professor, CSE, TKREC, Telangana, India [1]

Associate Professor, CSE, TKREC, Telangana, India [2]

CSE, TKREC, Telangana, India [3]

Abstract: Data mining emphases on producing beneficial information from the data sources rather than a simple data mining technology. Among several forms of data mining tasks, data classification is a key task for a user who wants to group a record at hand based on the database existing in the cloud. The existing privacy methods for privacy preservation in data mining are perturbation method and secure multiparty method (SMC). But these methods are current for the data that are not encrypted. Hence it is essential to present new protocols for the classification problem in data mining for the database stowed in cloud in encrypted form. We proposed the input record query classification problem of data mining task for accessing the database in encrypted form in the cloud is solved using KNN Classification method. A new privacy preserving protocol based on KNN classification method is used for protecting the privacy preservation and confidentiality of the data in the database, input query and data access.

Keywords: Security, KNN classification, Outsourced databases, Encryption.

I. INTRODUCTION

The advent of the cloud computing drives the organization to utilize the advantage of a cloud model. The cloud computing model provides benefits in cost, flexibility, administrative overheads etc. The cloud environment presents an opportunity for the data owner to outsource their database, database management functionalities and data mining tasks by providing the access mechanisms for querying and managing the hosted database. Though the data owner achieve the benefits in terms of cost and quality of service, the data mining tasks such as query processing and query hosting will be out of control data owner control leading to the challenges in privacy preserving on data and query as well as confidentiality on the data and query access patterns.

Encryption techniques provide a direct way to protect the confidentiality of the outsourced data from the cloud as well as from the unauthorized users by encrypting the data before outsourcing it. By this way, the data owner can protect the privacy of his own data. In addition, to preserve query privacy, authorized users require encrypting their queries before sending them to the cloud for evaluation. Furthermore, during query processing, the cloud can also derive useful and sensitive access patterns even if the data and query are encrypted. Hence the secure query processing needs to guarantee the confidentiality of the encrypted data, confidentiality of a user‟s query record and hiding data access patterns and these associated challenges are related to topic of privacy preserving data mining (PPDM).

The current PPDM methods like perturbation method and secure multiparty computation (SMC) cannot be applied in case of data that are in encrypted form. Since perturbed data does not have semantic security, perturbation techniques cannot be applied to encrypt the data. Usually the results of the perturbation methods are associated with statistical noises over the data and thus providing inaccurate data mining results. In case of SMC method the data are assumed to be distributed with each participating parties and are not encrypted. The most common type of input query processing in data mining is the classification type problem, where the user or authorized agents needs to classify his input record query based on the database that are outsourced to cloud. The input query record will be send to the cloud. The cloud computes the query record and provides the computed class label. Since the input query record will contain sensitive information it needs to be protected by way of encryption before sending it to cloud. If the database in the cloud is in encrypted form then the data mining task is called as data mining over encrypted data (DMED). In DMED data mining tasks, the user's record needs to be protected and the data access pattern also needs to be protected. Hence the privacy preservation problem in DMED data mining tasks are to preserve the confidentiality of the encrypted data in the database that is being outsourced to the cloud and to preserve the confidentiality of the user‟s input query record as well as to hide the data access pattern from the cloud. Apart from different classification methods that are available to solve such classification problem, K-NN

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

classification is the widely used industry methods for the application of privacy preservation scenarios and in domains such as similarity search, pattern recognition and data mining. Hence numerous Security protocols in data mining are being developed based on the K-NN methods for the privacy preservation in data mining applications. A novel security protocol, Privacy Preservation Protocol (PPKNN) based on K-NN classification method has been developed to meet the privacy preservation and confidentiality requirement in the DMED classification task. Considering data privacy, a traditional way to en-sure it is to rely on the server to enforce the access control after authentication [1], which means any unexpected privilege escalation will expose all data. In a shared-tenancy cloud computing environment, things become even worse. Data from different clients can be hosted on separate virtual machines (VMs) but reside on a single physical machine. Data sharing is an important functionality in cloud storage. For example, bloggers can let their friends view a subset of their private pictures; an enterprise may grant her employees access to a portion of sensitive data. Hence the secure query processing needs to guarantee the confidentiality of the encrypted data, confidentiality of a user‟s query record and hiding data access patterns and these associated challenges are related to topic of privacy preserving data mining (PPDM).The current PPDM methods like perturbation method and secure multiparty computation (SMC) cannot be applied in case of data that are in encrypted form. Since perturbed data does not have semantic security, perturbation techniques cannot be applied to encrypt the data. Usually the results of the perturbation methods are associated with statistical noises over the data and thus providing inaccurate data mining results. In case of SMC method the data are assumed to be distributed with each participating parties and are not encrypted [1].

The cloud computes the query record and provides the computed class label. Since the input query record will contain sensitive information it needs to be protected by way of encryption before sending it to cloud. If the database in the cloud is in encrypted form then the data mining task is called as data mining over encrypted data (DMED)[3]. In DMED data mining tasks, the user's record needs to be protected and the data access pattern also needs to be protected. Hence the privacy preservation problem in DMED data mining tasks are to preserve the confidentiality of the encrypted data in the database that is being outsourced to the cloud and to preserve the confidentiality of the user‟s input query record as well as to hide the data access pattern from the cloud. The k-nearest neighbors (k-NN) algorithm is a popular and effective classification algorithm. Due to its large storage and computational requirements, it is suitable for cloud outsourcing. However, k-NN is often run on sensitive data such as medical records, user images, or personal information [4]. It is important to protect the privacy of data in an outsourced k-NN system. Privacy preserving in k-nearest neighbor classifier is a simple data classifier, where the category (class label) of a data item is determined by using a majority vote of its neighbors, assigning the data item to the category most common among its k nearest neighbors. The input to the algorithm is a set of tuples, where one of the attributes is a class label and the other attributes are the features of the given data. While some of the tuples have known class labels, others have unknown class labels, and the task is to label those tuples with unknown class labels. The class label for any tuple t is determined with the following procedure:

- Measure the distance of the tuple t to each of the labelled tuples in the dataset.
- Find the setS of the k-nearest neighbors of t.
- Find the majority class label in the setS. If two or more classes have the majority decrease k by 1 until to find the majority class label.

II. RELATED WORK

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:

S. De Capitani di Vimercati, S. Foresti, and P. Samarati[1] Proposed a 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. The different aspects of the privacy are characterized in to privacy risk in the cloud, privacy risk for users, privacy risks for stored data and privacy risk for data access. P. Samarati and S. De Capitani di Vimercati [2] Approach Data outsourcing is an emerging paradigm that allows users and companies to give their (potentially sensitive) data to external servers that then become responsible for their storage, management, and dissemination. Although data outsourcing provides many benefits, especially for parties with limited resources for managing an ever more increasing amount of data, it introduces new privacy and security concerns. The main issues that needs to be addressed for guaranteeing proper protection and access to outsourced are, data protection, query execution, private access, data integrity and correctness, access control enforcement and private collaborative computation. A first solution used for preventing a server from accessing data stored on its own machines consists in encrypting the data before outsourcing them. H. Hu, J. Xu, C.Ren,and B. Choi [3] Proposed Query processing that preserves both the data privacy of the owner and the query privacy of the client is a new research problem. It shows increasing importance as cloud computing drives more businesses to outsource their data and querying services. However, most existing studies, including those on data outsourcing, address the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

data privacy and query privacy separately and cannot be applied to this problem. A holistic and efficient solution that comprises a secure traversal framework and an encryption scheme based on privacy homomorphism has been proposed. The framework is scalable to large datasets by leveraging an index-based approach. Based on that, a secure protocol for processing typical queries such as k-nearest-neighbor queries (kNN) on R-tree index has been proposed. P. Paillier [4] approach composite residuosity class problem is a significant computational methods applied to public-key cryptography. A new trapdoor permutation and two homomorphic probabilistic encryption schemes computationally comparable to RSA has been proposed. The proposed cryptosystems, based on usual modular arithmetic, are provably secure under appropriate assumptions in the standard model. The new trapdoor mechanism is based on composite residuosity in contrast to prime residuosity. The trapdoor provides a new cryptographic building-block for conceiving public-key cryptosystems. M. S. Islam, M. Kuzu, and M. Kantar cioglu [5] Implement Remote data storage offers reduced data management overhead for data owners in a cost effective manner. Sensitive documents, however, need to be stored in encrypted format due to security concerns. But, encrypted storage makes it difficult to search on the stored documents. Various protocols have been proposed for keyword search over encrypted data to address this issue. Most of the available protocols leak data access patterns due to efficiency reasons. A simple technique to mitigate the risk against the proposed attack at the expense of a slight increment in computational resources and communication cost was proposed.

The proposed mitigation technique is generic enough to be used in conjunction with any searchable encryption scheme that reveals data access pattern. The disclosure of data access patterns during „search over encrypted text". Pose a potential vulnerability. Formalized model that can be used to launch an inference attack utilizing this vulnerability and empirically show their efficiency in successfully predicting query identities. The „hiding access pattern" is extremely important in encrypted keyword search and therefore is a necessary characteristic of a secure encrypted search scheme. E. Shi, J. Bethencourt, T,-H, Chan, D. Song and A. Perrig[6] Proposed a scheme allows a network gateway to encrypt summaries of network flows before submitting it to an untrusted repository. When network intrusions are suspected, an authority can release a key to an auditor, allowing the auditor to decrypt flows whose attributes fall within specific ranges. However, the privacy of all irrelevant flows is still preserved. The security for Multi-dimensional Range Query over Encrypted Data and prove the security of our construction under the decision bilinear Diffie-Hellman and decision linear assumptions in certain bilinear groups was defined formally. Multi-dimensional Range Query over Encrypted Data implies a solution to its dual problem, which enables investors to trade stocks through a broker in a privacy-preserving manner. R. Agrawal and R. Srikant[7] Design a privacy preserving data mining was addressed for a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. The generic protocols in such a case are of no practical use and therefore more efficient protocols are required. X. Xiao, F. Li, and B. Yao [8] Proposed a mechanism to secure nearest neighbor problem, a client issues an encrypted query point to a cloud service provider and asks for an encrypted data point in the encrypted database that is closest to the query point, without allowing the server to learn the plaintexts of the data or the query and its result. The results imply that one cannot expect to find the exact encrypted nearest neighbor based on only encrypted query and encrypted data point. The proposed methods provide customizable trade-off between efficiency and communication cost, and they are as secure as the encryption scheme used to encrypt the query and the database.

Y. Qi and M. J. Atallah [9] Approach privacy preserving k-NN search is where two parties want to cooperatively compute the k nearest neighbors to a query point without revealing their private inputs to the other party. An improved the single-step protocol in both in terms of efficiency and privacy (no information leakage).It also gave a multi-step protocol for the algorithm proposed to meet the efficiency requirements of k-NN search with complex high- dimensional and adaptable distance functions and further gave a case study of sequence data as an application of our multi-step k-NN protocol. C. Gentry[10] proposed a totally homomorphic security arrangement is prescribed that permits one to survey circuits over secured data without having the capacity to decode. Our cure comes in three activities. Starting, we offer a typical result that, to assemble a security plan that permits appraisal of unessential circuits, it suffices to make a security plan that can survey (marginally upgraded releases of) its own unscrambling circuit; we contact an arrangement that can evaluate its (increased) decoding circuit boots trappable. Grid based cryptosystems for the most part have unscrambling calculations with low circuit multifaceted nature, frequently secured with an internal thing calculation that is in NC1.

III. PROPOSED SYSTEM

The proposed system focus on solving the classification problem over encrypted data. In the proposed system, a new privacy preservation protocol based on KNN classification method is introduced to protect the confidentiality of data, privacy of user"s input query and to hide the data access pattern. Figure 3.1 Architecture Diagram Using a Homomorphic encryption allows complex mathematical operations to be performed on encrypted data without using the original data and provides the data security in cloud. The proposed algorithm to preserve intermediate k nearest neighbor in the classification process should not revealed to cloud server or any other user. The proposed algorithm develops

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

a solution for privacy–preserving k-nearest neighbor classification which is one of the commonly used data mining tasks. It determines which the closest results are by identifying the class of minimum distance using K nearest neighbors. Refer figure 3.1 to privacy preservation for data in cloud.

The new privacy preservation protocol implementation for the input query record classification over the encrypted database in the cloud is carried by the steps,

• Secure Data Upload
• Query Processing
• Secure KNN query process

3.1 Secure Data Upload

The Admin have already to register itself. The admin is uploading the data to the cloud, before it should be encrypted. Two types of data uploaded into the cloud. One is the normal query process data (like voter list data), and the second one is kNN query process (like X and Y values). After that the user can be register to the cloud. The registered users only have permission for access the cloud data, so the unauthorized can‟t access the secure data from the cloud storage. Asymmetric –key algorithms require the use of asymmetric key pairs, consisting of a private key and corresponding public key. The key to be used for each operation depends on the cryptographic process being performed Each public / private key pair is associate with only one entity; this entity is known as the key-pair owner. The public key may be known by anyone, whereas the private key must be known and used only by the key-pair owner. Key pairs are generated by the key-pair owner

3.2 Homomorphic Encryption

The standard homomorphic encryption methods used for encrypting data. Homomorphic encryption is a form of encryption which allows specific types of computations to be carried out on cipher text and obtain an encrypted result which when decrypted matches the result of operations performed on the plaintext. An encryption scheme has algorithm consists of four steps: Homomorphic Encryption = {key generation, encryption, decryption, Evaluation}

Key generation: KeyGen is algorithm that generates pubik key, evaluation key and secret decryption key. The encrypted data, Using Homomorphic scheme, given two ciphertexts E(a) and E(b) of two plaintexts a and b respectively, an encryption of their sum E(a+b) can be efficiently computed by multiplying the ciphertexts modulo a public key n2, i.e., E(a +b) = E(a).E(b) mod n2.

 Encryption – Using secret key SK it encrypts the plaintext P and generate Esk (pt) and along with public key pubk this cipher text CT will be sent to the server.
 Decryption - decrypts the cipher text C with the privacy key privk to retrieve the plaintext P.
 Evaluation - outputs a cipher text C of f (P) such that Decrypt (privk, P) = f (P).

The scheme becomes Homomorphic if f can be any arbitrary function, and the resulting ciphertext of Evaluation is compact. That means it does not grow too large regardless of the complexity of function f. The Evaluation algorithm in essence means that the scheme can evaluate its own decryption algorithm. The Homomorphic Encryption has the best encryption method to ensure security and privacy of shared data.

3.3 Query Processing

After the user login to access the normal query window. In query process window, user to select the database name, table name and data owner access code from the database. In this process to protects the confidentiality of the data, user‟s input query, and hides the data access pattern. The user‟s input query will encrypted and pass to the cloud database. The cloud will classify label to corresponding query record. The query can retrieve the data from the cloud and show the encrypted and decrypted data in the output window.

3.4 Secure KNN query process

An authorized user sends the encrypted query to cloud server. The proposed PPKNN protocol is to classify user‟s query record using encrypted database in a privacy preserving manner. The PPKNN protocol has: PPKNN(Encrypted Database(D1),Query(Q))->Class Label(Cq). Where Cq denotes the class label for Q after applying k-NN classification method on D1 and Q. The KNN classification algorithm is a machine learning algorithm. It is a method for classifying objects based on closest training samples in the feature space. KNN is a type of instance-based learning ; many test records will not be classified because they do not exactly match any of the training records. A more sophisticated approach, k-nearest neighbor (kNN) classification, finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighbourhood.

There are three key elements of this approach: a set of labelled objects, e.g., a set of stored records, a distance or similarity metric to compute distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabelled object, the distance of this object to the labelled objects is computed, its k-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object. The proposed PPKNN protocol mainly consists of two stages:

Stage 1: Secure retrieval of K-nearest Neighbors (SRKNN). In this stage, the authorized user sends a query (in encrypted form) to cloud. After this cloud involve in a set of sub-protocols to securely retrieve(in encrypted form) the class labels corresponding to the K-nearest neighbors are known only to cloud server.

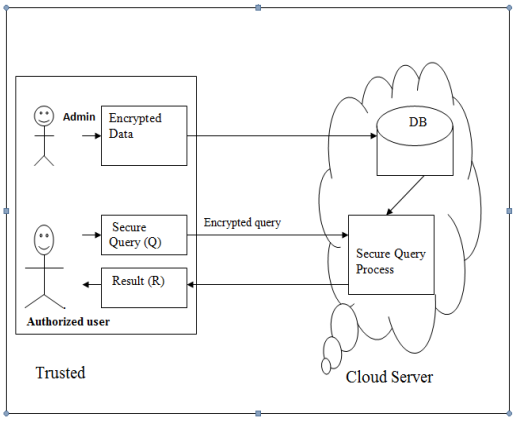
Stage 2: Secure Computation of Majority Class (SCMCk). Following from Stage 1, Cloud server will compute the

IJARCCE

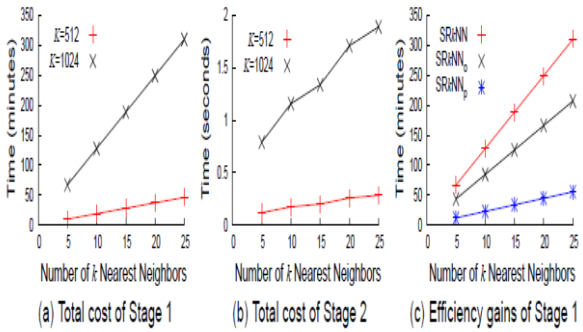
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

class label with a majority voting among the k-nearest neighbors of query. At the end of this step, only authorized



IV. EMPIRICAL EVALUATION



V. CONCLUSION AND FUTURE WORK

The new privacy preserving protocol based on KNN classification method is being applied to resolve the input classification difficulty based on the database that was outsourced to the cloud in the encrypted form. This protocol protects the privacy of the data, user's input query, and conceals the data access patterns. The Future Work focuses on the performance of the proposed protocol be contingent on the efficiency of the SMINn protocol. Improving the SMINn will be the first scope of future work. Implementing this new privacy preserving protocol algorithm in the other classification methods and comparing the performance of those classification methods with current KNN classification method will be the second scope of future work.

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] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Eurocrypt, pp. 223–238, 1999.
[5] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted relational data." Eprint arXiv:1403.5001, 2014.
[6] C. Gentry, "Fully homomorphic encryption using ideal lattices," in ACM STOC, pp. 169–178, 2009.
[7] C. Gentry and S. Halevi, "Implementing gentry's fullyhomomorphic encryption scheme," in EUROCRYPT, pp. 129–148, Springer, 2011.
[8] A. Shamir, "How to share a secret," Commun. ACM, vol. 22, pp. 612–613, Nov. 1979.
[9] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations," in ESORICS, pp. 192–206, Springer, 2008.
[10 R. Agrawal and R. Srikant, "Privacy-preserving data mining," in ACM Sigmod Record, vol. 29, pp. 439–450, ACM, 2000.
[11] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Advances in Cryptology (CRYPTO), pp. 36–54, Springer, 2000.
[12] P. Zhang, Y. Tong, S. Tang, and D. Yang, "Privacy preserving naive bayes classification," ADMA, pp. 744–752, 2005.
[13] 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.
[14] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in IEEE ICDE, pp. 217–228, 2005.
[15] 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.
[16] M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier," in PKDD, pp. 279–290, 2004.
[17] L. Xiong, S. Chitti, and L. Liu, "K nearest neighbor classification across multiple private databases," in CIKM, pp. 840–841, ACM, 2006.
[18] Y. Qi and M. J. Atallah, "Efficient privacy-preserving k-nearest neighbor search," in IEEE ICDCS, pp. 311–319, 2008.
[19] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in ACM SIGMOD, pp. 563– 574, 2004.
[20] 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.
[21] 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.
[22] 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.

BIOGRAPHIES

Dr V. Goutham is a Professor and Head of the Department of Computer Science and Engineering at Teegala Krishna Reddy Engineering College affiliated to J.N.T.U Hyderabad. He received Ph. d from Acharya Nagarjuna University and M.Tech from Andhra University. He worked for various MNC Companies in Software Testing and Quality as Senior Test Engineer. His research interests are Software Reliability Engineering, software testing, software Metrics, and cloud computing.

Mrs. P. Ashwini Reddy is working as a Assistant Professor in the Department of Computer Science and Engineering at Teegala Krishna Reddy Engineering College affiliated to J.N.T.U Hyderabad

Ms. K. Sunitha Department of Computer Science and Engineering at Teegala Krishna Reddy Engineering College affiliated to J.N.T.U Hyderabad.