K-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data

[1] K Sumalatha [2] Y AppaRao, [3] Prasad B
[1]II/IV, [2][3]Associate Professor
[1][2][3] Department of CSE, Marri Laxman Reddy Institute of Technology and Management (MLRITM) Hyderabad
[1] sumak9555@gmail.com [2] yapparao@gmail.com[3]bprasad@gmail.com

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 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. 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 proposed protocol using a real-world dataset under different parameter settings.

Keywords: **Data Mining, Classification, Mining Applications, Cloud Computing, Cloud Computing.**

I. INTRODUCTION

Recently, the cloud computing paradigm is revolutionizing N the organizations' way of operating their data particularly in the way they store, access and process data. As an emerging computing paradigm, cloud computing attracts many organizations to consider seriously regarding cloud potential in terms of its cost-efficiency, flexibility, and offload of administrative overhead. Most often, organizations delegate their computational operations in addition to their data to the cloud. 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. There are other privacy concerns, demonstrated by the following example .*Example 1:* Suppose an insurance company outsourced its encrypted customers database and relevant data mining tasks to a cloud. When an agent from the company wants to determine the risk level of a potential new customer, the agent can use a classification method to determine the risk level of the customer. First, the agent needs to generate a data record q for the customer containing certain personal information of the customer, e.g., credit score, age, marital status, etc. Then this record can be sent to the cloud, and the cloud will compute the class label for q. Nevertheless, since q contains sensitive information, to protect the customer's privacy, q should be encrypted before sending it to the cloud. The above example shows that data mining over encrypted data (denoted by DMED) on a cloud also 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 (either perturbation or secure multi-party computation 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 very accurate data mining results. Secure multi-party computation (SMC) based approach assumes data are distributed and not encrypted at each participating party. In addition, many intermediate computations are performed based on nonencrypted data. As a result, in this paper, 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.

II. SYSTEM ANALYSIS

2.1 Existing System

In the existing system, the system is implemented fully homomorphic cryptosystems can solve the DMED problem since it allows a third-party (that hosts the encrypted data) to execute arbitrary functions over encrypted data without ever decrypting them. However, we stress that such techniques are very expensive and their usage in practical applications have yet to be explored. For example, it was shown in the existing system that even for weak security parameters one “bootstrapping” operation of the homomorphic operation would take at least 30 seconds on a high performance machine.

2.2 Proposed System

In the proposed system, the system proposed novel methods to effectively solve the DMED problem assuming that the encrypted data are outsourced to a cloud. Specifically, the system focuses 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.

III. SYSTEM DESIGN

System Architecture:

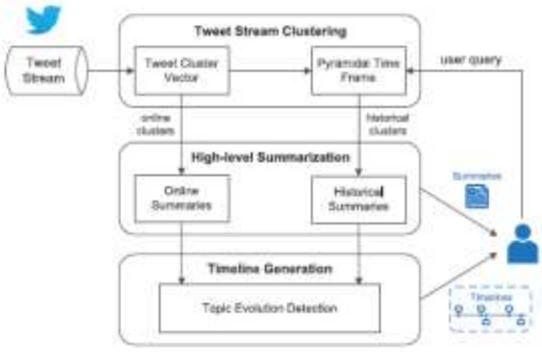

Fig 1: System Architecture

Data Flow Diagram:

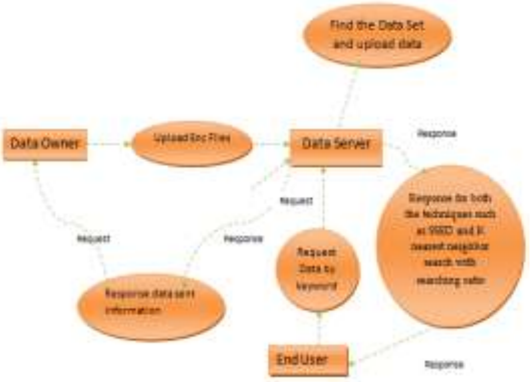
Find the Data Set and upload data
Data Owner
Upload Enc Files
Data Server
Response
Request
Request
Response
Response for both the techniques such as SSED and K nearest neighbor search with matching ratio
Request Data by keyword
Response data sent information
End User
Response

Fig 2 : Data Flow Diagram

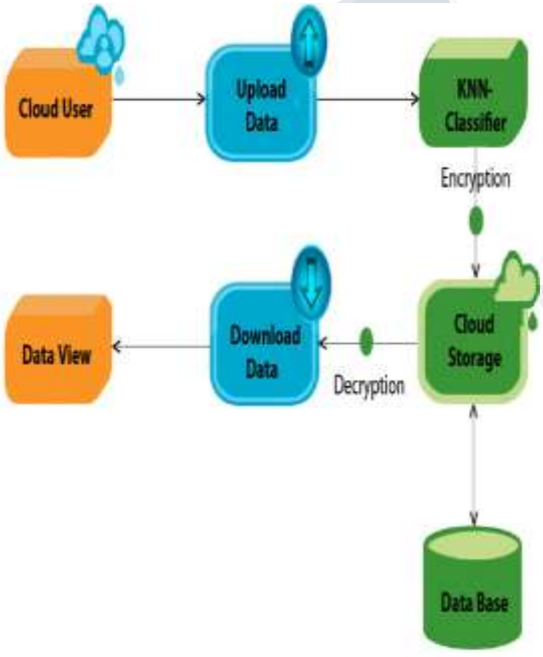

Fig 3 : System Design

Modules:

A. Data provider

In this module, the data provider uploads their data in the Data server. For the security purpose the data owner encrypts the data file and then store in the server. The Data owner can have capable of manipulating the encrypted data file.

B. Data Server

The Data server manages which is to provide data storage service for the Data Owners. Data owners encrypt their data files and store them in the Server for sharing with data consumers. To access the shared data files, data consumers download encrypted data files of their interest from the Server and then Server will decrypt them. The server will generate the aggregate key if the end user requests multiple files at the same time to access.

C. End User

In this module, the user can only access the data file with the encrypted key word. The user can search the file for both the methods such as SSED and K nearest neighbor search. The user has to register and then login from the Data server.

IV. IMPLEMENTATION

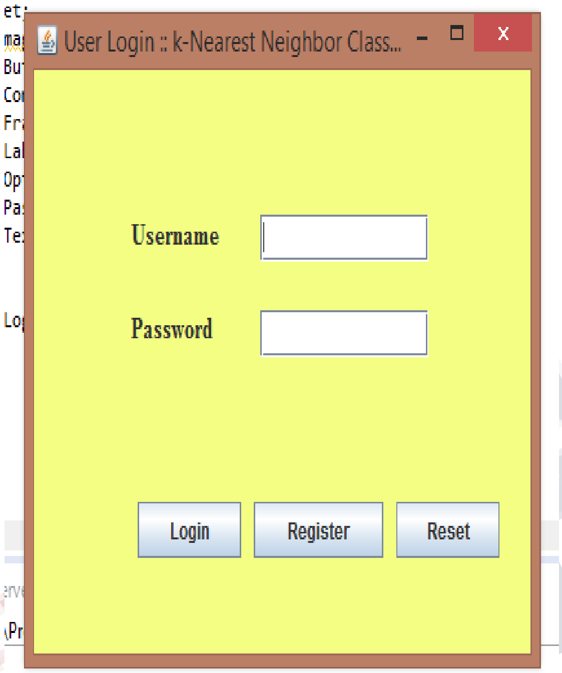

Fig 4: User Login

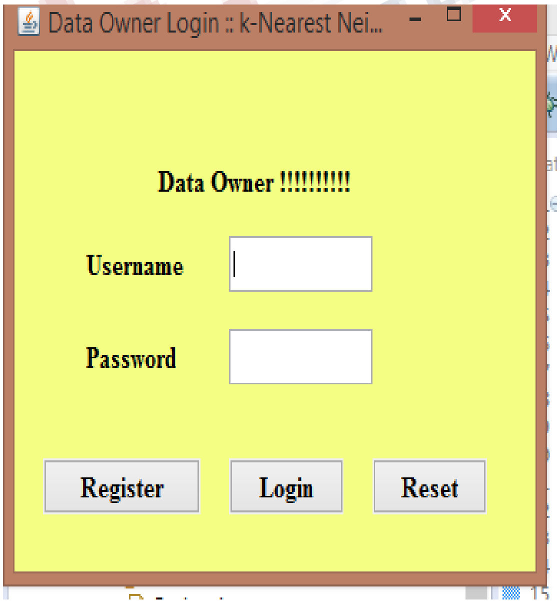

Fig 5: Data Owner Login

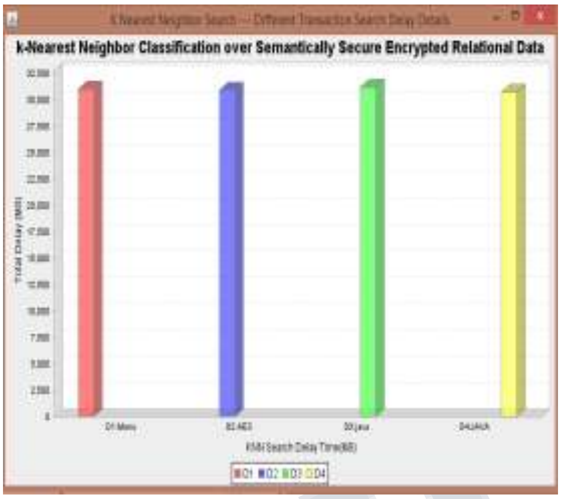

Fig 6: Nearest Neighbour Search

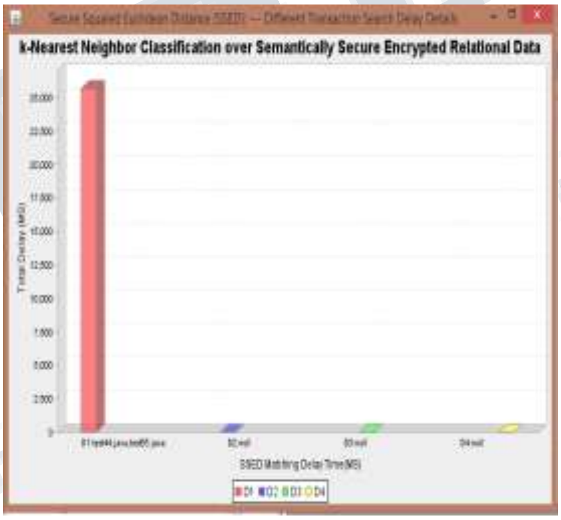

Fig 7: Search Squared Euclidean Distance

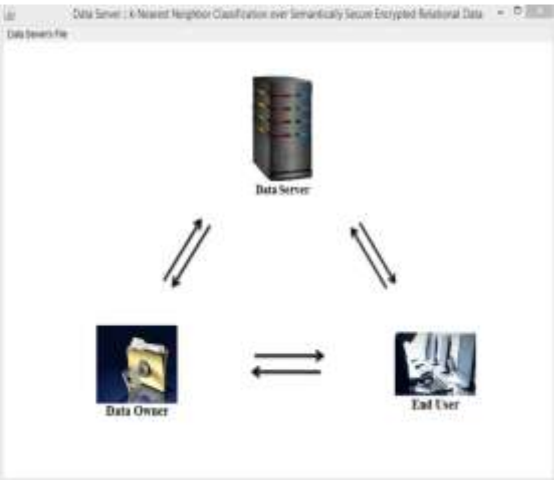
Data Server :: k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data
Data Server's File
Data Server
Data Owner
End User

Fig 8: Relational Data

V. CONCLUSION

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. Since improving the efficiency of SMINn is an important first step for improving the performance of our PPkNN protocol, we plan to investigate alternative and more efficient solutions to the SMINn 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 *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 compositedegree 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.