International Journal of Computer Applications (0975 - 8887)
Volume 133 - No.8, January 2016

Review on k -Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data

Swathy J.
P G Scholar
Department of Computer Science & Engineering
College of Engineering, Perumon(CUSAT), Kerala, India

Surya S.R.
Assistant Professor in CSE
Department of Computer Science & Engineering
College of Engineering, Perumon(CUSAT), Kerala, India

ABSTRACT

Data mining has wide variety of real time application in many fields such as financial, telecommunication, biological, and among government agencies. Classification is the one of the main task in data mining. For the past few years, due to the increment in various privacy problem, many conceptual and feasible solution to the classification problem have been proposed under different certainty prototype. With the increment of cloud computing users have an opportunity to offload the data and processing to the cloud, in an encrypted form. The data in the cloud are in encrypted form, existing privacy preserving classification systems are not relevant. This paper reviews how to perform privacy preserving k-NN classification over encrypted data in the cloud. The recommended protocol preserves privacy of data, protect the user query, and hide the access mode.

General Terms

Data Mining, Classification

Keywords

Security, k-NN classifier, outsourced databases, encryption

1. INTRODUCTION

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. Data mining sometimes called data or knowledge mining. Data are may be numbers, or sequence of characters that can be processed by a computer.

Now a days cloud computing model [1] is changing the structure of the organizations way of storing, accessing, and processing their data. As the growing processing data, many organizations to focus on the cloud computing in terms of its efficiency, flexibility, security, and document control .To reduce the data overhead the companies offload their data to the cloud. Most often, organization give their computational activity in addition to their data to the cloud. For all massive advantages of cloud computing provide, privacy and security problems in the cloud are blocking organization to use this advantages. The data are in encoded form, traditional encryption scheme execute any data mining function over encrypted data is very challenging process. There are other privacy problems shows by the following example.

Example: Assume an insurance company offload its encoded customers data set and appropriate data mining task to the cloud .Insurance company representative to find the risk level of newly arriving customer. The representative uses the classification method to find the risk level of the customer. First the representative wants to create data record q for the customer containing the personal details like age, sex, customer id, etc. Then this data can be send to the cloud, and the cloud will calculate the class label for q. Since q hold sensitive details to protect the customer privacy q should be encrypted before outsource to the cloud.

The above example reveals that the Data Mining over Encrypted Data (denoted by DMED) cloud also want to protect the users data when the data is the part of a data mining process. Also cloud can prove useful and sensitive data about the original data items by detect data access patterns even if the data are in encrypted format. For the privacy and security reasons DMED issue have main three reasons (1) privacy of the encrypted data (2) security of the users query record and, (3) conceal data access pattern.

Current work on privacy preserving data mining cannot solve the DMED issues. Secure multi-party computation model do not provide semantic security therefore cannot use this method for encrypted data. Also the secure multi-party computation model the data is distributed but each party cannot encrypt the data. Therefore the number of intermediate operations is performed over the non-encrypted data. As a result this protocol solves the DMED problem effectively by offload the encrypted data in to the cloud. The architecture of this system is shown in Fig. 1

The classification is the main task in data mining. The privacy preserving data mining operations (Classification/Clustering etc.) has thus become an important problem in current years. Each classification has its own advantages, is provide privacy preserving k-Nearest Neighbor classification over encrypted data in the cloud computing environment is a challenging problem.

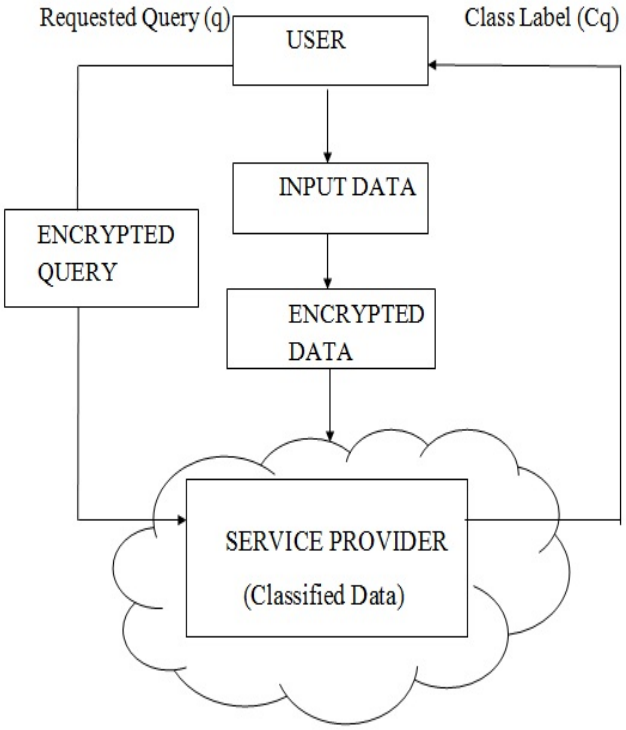

Fig. 1. Architecture of k-NN classification over semantically secured encrypted relational data

2. LITERATURE SURVEY

In the past, traditional encryption schemes were used for data security. Here the client sends both the cipher text and the private key to the server. The server then decrypts and processes them by using the client's private key. If the client's private key is compromised, any intruder can easily access the sensitive data [2]. This encryption is not applicable in cloud environments, as the cloud is an open structure, so that any number of intruders can gain access to the private data. As a result, the traditional methods fail in providing a better security. In 1979, Shamir introduced the first secret sharing scheme [3] consisting of following steps,

(1) Choose any prime number $pmax(s, n+1)$. Let Zp represents the field of integers modulo p.
(2) Choose $a1, a2, ...ak-1 \epsilon Zp$, randomly, uniformly, independently.
(3) Let $q(x) = D+a1x+a2x2+.....+ak-1xk-1$
(4) Let $Di = q(i), \forall 1 \leq i \leq n$. (The evaluation of $q(i)$ done over Zp)

This scheme was not secure against intruders. So in 1998, Thomas J introduced an efficient and accurate secret sharing method [4]. The method worked on the following basis: Assume that probability of undetected cheating is less than ε, for any $\varepsilon > 0$.

(1) Choose any prime number $pmax((s-1), (k-1)/\varepsilon + k, n)$.
(2) Choose $a1, a2, ...ak-1 \epsilon Zp$, randomly, uniformly, independently.
(3) Let $q(x) = D+a1x+a2x2+.....+ak-1xk-1$
(4) 4. Choose $(x1, x2, ...xn)$ uniformly and randomly from among all permutations of all distinct elements from $1, 2, 3.....p-1$. Let $Di = (xi, di)$, where $di = q(xi)$

Suppose, the $k-1$ participants are intruders, these methods cannot handle the existing security problems. The marten van Dijk and Ari Juels are prove traditional encryption schemes are not solve the privacy preserving problem in cloud computing environment [5]. In 2009, Craig Gentry introduced a Fully Homomorphic encryption scheme [6] aiming for a better level of security. The scheme could evaluate the circuit over encrypted data without employing any decryption process. The system works using the following steps:

(1) Using encryption to evaluate the arbitrary circuit.
(2) Using same encryption to find its own decryption circuit.
(3) Using encryption scheme to find its decryption circuit boots trappable.

Even though the method is more secure than normal encryption schemes, it is very expensive and their usage in real time applications has not yet been explored fully. Homomorphic encryption serves as a cheap alternative which makes it an apt choice to be used in various real world applications. Homomorphic encryption is to provide security for large data. Current methods for privacy preserving data mining [7] cannot be applied to Data Mining over Encrypted Data problems (DMED). This problem can be solved using a Secure Multi-party computation model [8]. In this method, sensitive data is collected and processed and also they provide an efficient general purpose computation system to address this issue. Share mind is a virtual machine for privacy-preserving data processing that depends on the share computing method.

Nowadays,there exist different methods of classification techniques [9] in data mining. Each classification technique has its own merits and demerits. In 2006, Murat Ksantarcioglu and Chris Clifton introduced privately computing k-nearest neighbor classification in a distributed manner [10]. If any user want to find the class label for an attribute x, the user sends a query to the each of the systems so that each system individually calculates the k-nn based on Euclidean distance between attributes and x. After processing each user sends the partial result to non colluding third party. The outsourced partial result is encrypted using the user's public key. Therefore the third party cannot access the result. Third party to calculate the final class label based on the partial result. User accepts the final result from the non-colluding third party and decrypts it. This method use the normal encryption scheme for security, there for the data are not that much secure. The architecture of this system is shown in Fig. 2

Provide privacy preserving k-nearest neighbour classification [11] is to use the semantically secure homomorphic encryption scheme. The privacy preserving k-nearest neighbour classification technique can classify large amounts of data in a much secured fashion. In this technique, user gives a query xq wants to compute the distance between xq and each training instances. Each user handle only portion of the instances they compute the distance measure according to their attribute values. Find the k-nearest neighbour of xq all users to sum their distances together. To obtain the resulting distance without losing data privacy,one user to

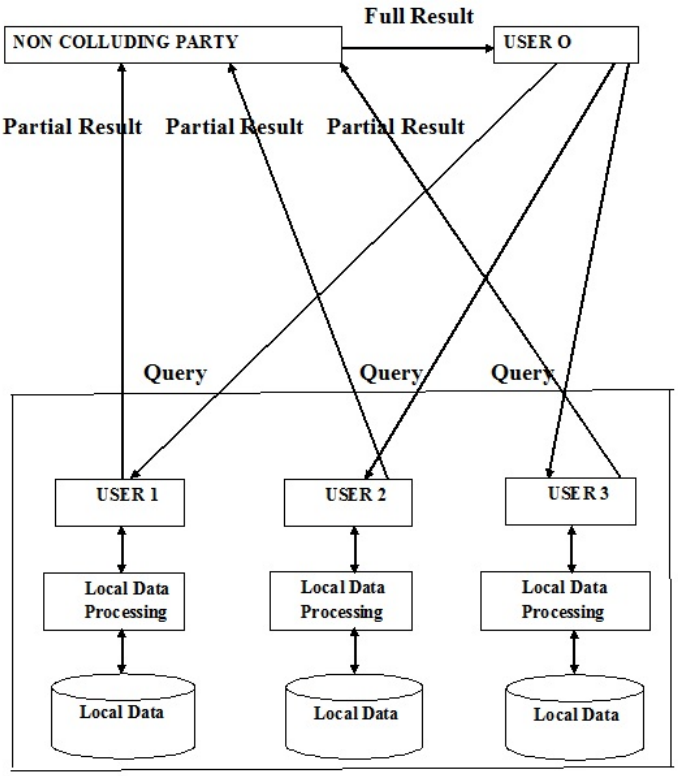

Fig. 2. Information flow in secure k-nn classifications

generate multiple queries and collect the distance (Euclidian) from each query and to find the final result.

3. EXPERIMENTAL RESULTS

Here discussed some experiments demonstrating the performance of Privacy Preserving k-Nearest Neighbor(PPKNN) classification method with some parameter settings. The Partial Homomorphic encryption scheme is used as the underlying additive homomorphic encryption scheme and implemented the proposed PPkNN protocol in JAVA.

3.1 Dataset Details and Experimental Setup

This protocol used the Car Evaluation dataset from the UCI KDD archive[12]. This dataset consists of 1728 records (n = 1728) and 6 attributes (m = 6). There is a separate class attribute and the dataset is categorized into four different classes(w = 4). Encrypted this dataset attributewise, using the Homomorphic encryption the key size is varied in experiments, and the encrypted data were stored on server machine. Based on PPkNN protocol, executed a random query over this encrypted data.

3.2 Performance of Privacy Preserving k-Nearest Neighbor Classification with Partial Homomorphic Encryption

The encryption key size K is either 512 or 1024 bits if is K=512 bits, the computation cost varies from 9.98 to 46.16 minutes when k is changed from 5 to 25, respectively. When K=1024 bits, the computation cost varies from 66.97 to 309.98 minutes when k is changed from 5 to 25, respectively. For K=512 bits, the computation time for Stage 2 to generate the final class label corresponding to the input query varies from 0.118 to 0.285 seconds when k is changed from 5 to 25. For K=1024 bits, Stage 2 took 0.789 and 1.89 seconds when k = 5 and 25, respectively. Here observed that the computation time of Stage 1 accounts for at least 99 percentage of the total time in PPkNN. For example, when k = 10 and K=512 bits, the processing costs of Stage 1 and 2 are 19.06 minutes and 0.175 seconds, respectively. Under this scenario, cost of Stage 1 is 99.98 percentage of the total cost of PPkNN. The total computation time of PPkNN grows almost linearly with n and k.

4. 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

5. CONCLUSION

Different type of privacy preserving classification has been introduced in past few years. This method is not applicable to outsourced databases. This protocol proposed an efficient privacy preserving k-NN classification over encrypted data in the cloud and is used to preserve privacy of the data, security of user query, and hide the access pattern. Efficiently perform a classification over encrypted data using the privacy preserving k-nearest neighbor(PPKNN) classification. The future work is to perform searching over encrypted documents in the cloud. Use this PPKNN classification for document classification and documents are encrypted using fully homomorphic encryption.

6. REFERENCES

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

[2] Subodh Gangan , "A Review of Man-in-the-Middle Attacks".

[3] A. Shamire, "How to share a secret" Commun. ACM, vol. 22, pp. 612-613, Nov. 1979.

[4] IBM Thomas J, "How to share secret with cheaters" Journal of Cryptology, I:133-138, 1988.

[5] Martenvan Dijk and Ari Juels , "On the Impossibility of Cryptography alone for Privacy-Preserving Cloud computing".

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

[7] R.Natarajan1, Dr.R.Sugumar, M.Mahendran and K.Anbazhagan, " A survey on Privacy Preserving Data Mining " International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 1, March 2012.

[8] Y. Lindell and B. Pinkas, "Privacy preserving data mining" in Advances in Cryptology (CRYPTO), pp. 3654, Springer, 2000.

[9] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining " Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol IIMECS 2009, March 18 - 20, 2009, Hong Kong.

[10] M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier" in PKDD, pp. 279290, 2004.

[11] Justin Zhan,LiWu Chang and Stan Matwin, "Privacy Perserbing K-nearest Neighbor Classification" International Journal of Network Security ,Vol.1 No,1,PP.46-51,July 2005.

[12] M. Bohanec and B. Zupan, "The UCI KDD Archive" http: //archive.ics.uci.edu/ml/datasets/Car+Evaluation, 1997.