Fast Search Processing Over Encrypted Relational Data Using K-Nearest Neighbour Algorithm

Bhagyashree Ambulkar
M. Tech, Department of Comp. Sci. and Engineering
Nagpur Institute of Technology, Nagpur
Nagpur, Maharashtra
bhagya.ambulkar@gmail.com

Prof. Gunjan Agre
M. Tech,Department of Comp. Sci. and Engineering
Nagpur Institute of Technology, Nagpur
Nagpur, Maharashtra
gunjan.agre@gmail.com

Abstract — Data mining has been used in real time application in a number of areas such as for example financial, telecommunication, biological, and among government agencies and several application handle very sensitive data. So these data remains secure and private.Data encryption is a very strong option to secure the data in databases from unauthorized access and intruder.The previous privacy preserving classification techniques are not feasible for encrypted data of database.In this paper, our proposed method provides privacy-preserving classifier for encrypted data of relational databasesand achieves the better performance for extracting information from encrypted data of relational databases.

Keywords- *cloud, Encryption, Privacy Preserving Classification, kNN classification, database encryption*

*****

I. INTRODUCTION

For the past decade, query processing on relational data has been studied widely, and many theoretical and practical solutions to query processing have been proposed under various scenarios. With the recent utilization of cloud computing, users are able to outsource their data in addition to the data management tasks to the cloud. However, as a result of rise of varied privacy problem, sensitive data e.g., transactional data have to be encrypted before outsourcing to the cloud. Additionally, query processing tasks should really be handled by the cloud; otherwise, there could be no point out outsource the data at the first place. To process queries over encrypted data with no cloud ever decrypting the data is really a very challenging task. Information industry has huge amount of data. For providing privacy and security for this data encryption techniques are applied. When user extracts useful information from it, the searching performance degraded.

II. RELATED WORK

Bharath K. Samanthulaet. al. [1] 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 provides the security 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. Author also, analyze the efficiency of proposed protocol using a real-world dataset under different parameter settings.[12]

E.Vaniet. al. [2] proposed to protect user privacy, various privacy-preserving classification techniques. The existing techniques are not applicable to outsourced database environments where the data resides in encrypted form on a third-party server. A novel privacy preserving k-NN classification protocol over semantically secure relational encrypted data in the cloud is proposed. The proposed algorithm to preserve intermediate k nearest neighbor in the classification process should not reveal to cloud server or any other user. The proposed algorithm develops a solution for privacy preserving k-nearest neighbor classification which is one of the areas used in data mining tasks. It determines which the closest results are by identifying the class of minimum distance using K nearest neighbors[12].

Zhihua Xia et. al. [3] present a secure multi-keyword ranked search scheme over encrypted cloud data, which simultaneously supports dynamic update operations like deletion and insertion of documents. Specifically, the vector space model and the widely-used TF × IDF model are integrated in the index construction and query generation. Author constructs a special tree-based index structure and proposes a "Greedy Depth-first Search" algorithm to provide efficient multi-keyword ranked search. The secure kNN algorithm is utilized to encrypt the index and query vectors, and meanwhile ensure accurate relevance score calculation between encrypted index and query vectors. In

order to resist statistical attacks, phantom terms are added to the index vector for blinding search results. Due to the use of proposed tree-based index structure, the proposed technique can achieve sub-linear search time and deal with the deletion and insertion of documents flexibly [12].

YousefElmehdwiet. al. [4] 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. Author first present a basic scheme and demonstrate that such a naive solution is not secure. To provide better security and proposed a secure kNN protocol that provides the better confidentiality of the data, user's input query, and data access patterns. Author also empirically analyzes the efficiency of our protocols through various experiments. The result shows that proposed 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 [12].

Hong Ronget. al. [5] focuses on privacy-preserving k-Nearest Neighbor (kNN) computation over the databases distributed among multiple cloud environments. Unfortunately, existing secure outsourcing protocols are either restricted to a single key setting or quite inefficient because of frequent client-to-server interactions, making it impractical for wide application. To address these issues, author proposed a set of secure building blocks and Outsourced Collaborative kNN (OCkNN) protocol. Theoretical analysis shows that proposed scheme not only preserves the privacy of distributed databases and kNN query, but also hides access patterns in the semi-honest model. Experimental evaluation demonstrates its significant efficiency improvements compared with existing methods [12].

Jianglin Huang et. al. [6] study the novel grey relational analysis based incomplete-instance kNN imputation is built for software quality data. An evaluation is conducted on four quality datasets with different simulated missingness scenarios to analyze the performance of the proposed imputation. The empirical results show that the proposed approach is superior to traditional kNN imputation and mean imputation in most cases. Moreover, the classification accuracy can be maintained or even improved by using this approach in classification tasks [12].

III. Classification Algorithm

In the original k-nearest neighbor (KNN) classification technique, classifier model is not built in advance. The whole training set is considered as a classifier. The simple idea is that the most similar tuples most belongs to the same class(a continuity assumption). Based on some pre-defined distance, the k nearest training samples of the sample to be classified and allocate the verity of classesof those k samples to the new sample data. For k, the value is pre-selected. If value of k is relatively large then k may not include similar pixels and on the other hand, if the value of k is small then k may exclude some potential candidate pixels.In both cases the classification accurateness will reduce. The optimal value of k depends on the size and nature of the data.

The usual value for k is 3, 5 or 7. The classification process steps are as follows:

1. Define a suitable distance measure.
2. Find the closed k nearest neighbors in training datasetusing the selected distance measure.
3. Find the number of class labels of the k-nearest neighbors
4. Assign that class label to the sample data which to be classified.

In this paper we provide a knn algorithm using RTree based on Euclidean distance measure to improve the performance of searching the nearest neighbors.Instead of examining individual pixel to find the nearest neighbors, to start initial neighborhood with the target sample and then consecutively expand the neighborhood area until there are k pixels in the neighborhood set. Neighborhood is a set of neighbors of the target pixel within a specified distance based on Euclidian distance measure.The expansion of neighborhood is done in such a way thatit alwayscovers the closest or most similar pixels of the target sample. There may be possible that the more boundary neighbors are present in equidistance form sample and which are required essentially to complete the k nearest neighbor set. In such situation, one can either use the larger set or randomly ignore some of them.

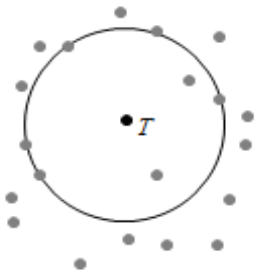

Figure 1: T, the pixel in the center is the target pixels. With k = 3, to find the third nearest neighbor, there are five pixels on the boundary line of the neighbourhood which are equidistant from the target.

Instead of proposing a novel approach of constructing nearest neighbor (NN) set, where we take the closure of the k-NN set which include all of the boundary neighbors and we call it the best-KNN set.Best-kNN is a superset of kNN set.

In the above example, with the values of k = 3, kNN includes the two points inside the circle and any one point on the boundary. The best-KNN includes the two points

inside the circle and all of the five boundary pixels.The algorithm of the best-kNN set is given below.

***Algorithm*:**

1. *if a ∈kNN, then a ∈ closed-kNN*
2. *if a ∈best-kNN and d(T,b) ≤ d(T,a), then b ∈ closed-KNN*
   Where, d(T,a) is the distance of a from target T.
3. *best-kNN does not contain any pixel, which cannot be produced by step 1 and 2.*

The best-kNN produces higher classification accurateness than KNN does.The closed-kNN encompassedonly those pixels, which are in as equal distance as some other pixels in the neighborhood without further expanding the neighborhood. In hybrid kNN, extra computation is not compulsory to find the closed-kNN. The nearest neighborhood automatically includes the points on the boundary of the neighborhood.

IV. PROPOSED SYSTEM

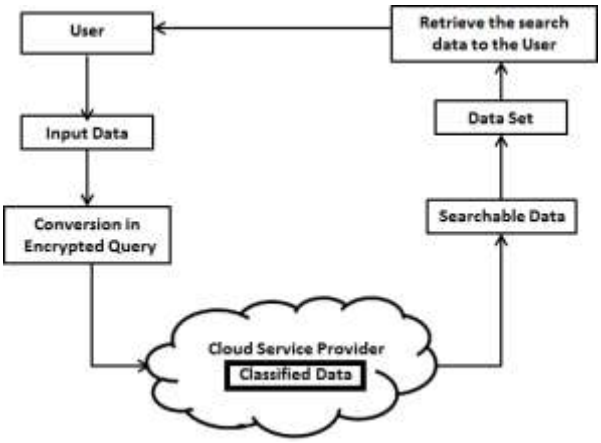

Figure 2. Proposed System Design

Existing methods on Privacy preserving data mining cannot resolve the Data Mining on Encrypted Data problem. The traditional search process on encrypted data is performed by decrypting the whole data and then retrieves the data. This whole task takes so much time and also reduces the performance of searching [7]. The proposed method provides good solution for privacy preserving and fast data retrieval.

The proposed method consists of two stages. In first stage, classification is performed over encrypted data to create the class labels and second stage executes searching task over classified data using RTree algorithms. Steps for propose method is as follows:

1. Upload dataset
2. Apply the encryption on data
3. Store encrypted data on cloud
4. Apply best-kNN classification on encrypted data which creates class labels
5. Fast retrieval of data using RTree
6. Display result.

A. Finding best- kNN set:

Input: P_i for all i, all the row id of training dataset and target row id.

Output: best- kNN set

For i= 1 to n-1 do

$R_{id} \leftarrow P_{ri}(V_i)$

$R_{id} \leftarrow P_{r1}$

For i=2 to n-1 do

$R_{id} \leftarrow R_{id} \& P_{rn}$

B. Searching object using RTree:

Input: classified dataset

Output: the value of RTree

We can apply the searching of an R-tree to find objects that overlap a search object, say obj, by the following steps.

SearchObject (T, obj)

1. if T is not a leaf node
2. search object obj in bounding box
3. for each entry e in T call SearchObject (T, obj)
4. else
5. search all entries e and find those bounding box which intersect to e
6. add these entries to the answer dataset
7. end

First of all the given dataset is classified according to user query requirement. The elements which are having the similar properties and label values get group together to form separate data set. Then remaining dataset processes according to the next user query [9].

After classification all data of user transformed into other data sets and on this dataset we apply RTree algorithm if the condition in RTree algorithm is tree. It finds the root element of the tree and then find leftmost and rightmost element of the tree. If data on left side match with kNN value it find the next left and right child of current root element. In this pattern search all elements [9].

V. CONCLUSION

Various types of privacy preserving classification have been introduced in past few years. These methods are not applicable to outsourced databases. Here we proposed a method to improve performance of data classification and searching over cloud encrypted data. The proposed method makes the system highly scalable and minimizes information leakage. It prevents overload by ranking the

files at the user side, reducing bandwidth and protects document frequency. The proposed solution is secure, scalable, accurate and fast compared to the other ranked keyword search.

REFERENCES

[1] Bharath K. Samanthula,Yousef Elmehdwi,Wei Jiang,"k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 5, May 2015.

[2] E.Vani, S.Veena, D.John Aravindar, "Query Processing Using Privacy Preserving k-NN Classification Over Encrypted Data", International Conference On Information Communication And Embedded System (ICICES), 978-1-5090-2552-7, 2016.

[3] Zhihua Xia, Xinhui Wang, Xingming Sun, Qian Wang,"A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data",IEEE Transactions On Parallel And Distributed Systems, Vol. 27, No. 2, February 2016.

[4] Yousef Elmehdwi, Bharath K. Samanthula, Wei Jiang, "Secure k-Nearest Neighbor Query over Encrypted Data in Outsourced", ICDE IEEE Conference,978-1-4799-2555-1/14/$31.00 © 2014.

[5] Hong Rong, Huimei Wang, Jian Liu, and Ming Xian, "Privacy-Preserving k-Nearest Neighbor Computation in Multiple Cloud Environments",IEEE, 2169-3536 (c) 2016.

[6] Jianglin Huang, Hongyi Sun, "Grey Relational Analysis based k Nearest Neighbor Missing Data Imputation for Software Quality Datasets", IEEE, International Conference on Software Quality, Reliability and Security, 2016.

[7] Manish Sharma , Atul Chaudhary, Santosh Kumar, "Query Processing Performance and Searching over Encrypted Data by using an Efficient Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 62– No.10, January 2013

[8] Kavyashree J, Deepika N, "Secured Access to Cloud Data through Encryption and Top-K Retrieval Using Multiple Keywords", International Journal of Science and Research (IJSR) ISSN (Online): 2319-706

[9] Maleq Khan, Qin Ding and William Perrizo, "K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees".

[10] http://www.bowdoin.edu/~ltoma/teaching/cs340/spring08/Papers/Rtree-chap1.pdf

[11] Jinka Sravana , Suba. S, "Applying R Trees In Non Spatial Multidimensional Databases", International Journal Of Technology Enhancements And Emerging Engineering Research, Vol 2, Issue 7 28 ISSN 2347-4289

[12] Bhagyashree Ambulkar, Prof. Gunjan Agre, "Review on Secure Encrypted Relational Data by Knn Classification Approach", Special Issue Tech-Ed 2017, 3rd Annual International Conferenceon Encouraging Innovation and EntrepreEngineering Education , Global Partnership for Best Practices, Policy Framework for Governance & Quality Enhancement organized by TGPCET, Nagpur, ISSN NO: 2454-1958