IMPROVISATION OF K-NN CLASSIFIER ON SEMANTICALLY SECURE ENCRYPTED RELATIONAL DATA

K. Sowjanya
M. Tech Scholar, Department of CSE,
Sree Vidyanikethan Engineering College (Autonomous),
Sree Sainath Nagar, Tirupati, India

A. V. Sriharsha
Associate Professor, Department of CSE
Sree Vidyanikethan Engineering College (Autonomous),
Sree Sainath Nagar, Tirupati, India

ABSTRACT

By the rapid improvement in web help and their popularity, web customers are developing day by day. Hence, there is large and various data. Data Mining has a wide use for the fields of business, medicine, experimental research and among government offices. One of the generally used tasks in data mining applications is Classification. Various professional and possible solutions to the classification problem have occurred introduced in the earlier decades. To overcome the privacy problems, certain clarifications have used various security types. Customers can outsource their data onto encrypted information and the data mining tasks to the cloud. Current privacy preserving classification techniques are not suitable for this encrypted data over the cloud. Securing proper privacy and protection of the data stored, transmitted, prepared, and distributed among the cloud as well as from the users entering such data is one of the big challenges to our current society. Hence, this paper proposes to define the classification problem of encrypted data. A k-NN classifier over encrypted data onto the cloud is proposed here for security reason. This method proposes a protocol to implement the confidentiality of the data onto the cloud protects the privacy of user input query and hides the data access patterns of the cloud. A certain k-NN classifier is the first above the encrypted data onto the semi-honest form. Since developing the performance of SMINn is an essential first step for improving the performance of our PP-k-NN protocol, the alternative and more effective solution than SMINn is studied that extends to different classification algorithms.

Key words: Security, k-nn Classifier, Outsourced Databases, Encryption
1. INTRODUCTION

1.1. K-nn Classification

K- Nearest neighbors are single algorithms that collect all possible cases and classifies various states based on a comparison measure (e.g., distance functions). K-NN has been used in mathematical estimation and model classification previously into the start of 1970’s as a non-parametric method.

K-NN can be used for both classification and regression predictive problems. However, this is also generally used in classification problems of the business. To evaluate any method we normally look at 3 important characters:

1. Ease to perform output
2. Prediction times
3. Predictive Power

The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, this input consists of the k- nearest training examples of the feature location. The output depends on whether k-NN is used for classification or regression.

Ink-NN classification, the output is a class membership. An object is selected by a majority voting for its neighbors, by the object moving allotted to the class very familiar amongst its k-nearest neighbors (k is a positive integer, typically small). If k = 1, if the object is generally selected to the class of that particular nearest neighbor.

Ink-NN regression, the output is the business price for the object. Each value is the average of the values of its k nearest neighbors.

K-NN is a type of instance-based learning, or lazy learning, wherever the function is simply approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, that can be important to allow the weight to the participation in the neighbors, hence that the nearer neighbors add extra to the normal than the more different ones. For example, a simple weighting design consists in providing every neighbor a weight of 1/d, where d is the distance between the neighbors.

These neighbors are selected from a collection of objects for what the class (fork-NN classification) or the object attribute value (fork-NN regression) is identified. That can be thought of as the training set for the algorithm if no explicit training step is needed. A shortcoming of the k-NN algorithm is that it is delicate to the local building of the data. This algorithm is not to be included ink-means, several modern tools learning method. The training examples are vectors in a multidimensional feature location, individual with a class label. Each training aspect of the algorithm consists only of collecting the feature vectors and class labels of the training samples. In the classification aspect, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label that is most frequent amongst the k training examples nearest to that query point.
A generally accepted distance metric for constant variables is Euclidean distance. Since discrete variables, such as for text classification, the various metric can be implemented, such as the extension metric (or Hamming distance). In the meaning of gene expression microarray data, for example, k-NN has also been used to change coefficients such as Pearson and Spearman. Normally, the classification performance of k-NN can be improved significantly if the distance metric is defined by particular algorithms such as Large Margin Nearest Neighbor or Neighborhood elements analysis. A disadvantage of the basic “majority voting” classification occurs to the class distribution is skewed.

This is, examples of a more common class tend to manage the prediction of this new example because they manage to be accepted with the k nearest neighbors due to their huge number. One approach to overcome the problem is to weight the classification, managing into the record the distance between the test points of various of its k nearest neighbors. This class (or value, in regression problems) of each of the k the nearest point is generated by a weight similar of the inverse of the distance between the position of the test point. A different approach to overcome to skew is by abstraction in data representation. For example, in a self-organizing map (SOM), every connection is a representative (a center) of a cluster of similar points, regardless of their density in the initial training data. K-NN can then be implemented to the SOM.

An example of k-NN classification. This analysis sample (green circle) should be classified both to the original class from blue rectangles or to the second class of red triangles. If k = 3 (solid line circle) it is elected to the second class since there are 2 triangles and only 1 square in the central circle. If k = 5 (dashed line circle) it is allotted to the first class (3 squares vs. 2 triangles inside the outside circle. The k-nn classification is shown in fig 1.1.

2. RELATEDWORK

2.1. A General Review of Privacy-Preserving Data Mining Patterns and Algorithms
Charu C. Aggarwal and Philip S. Yu described that a review of this state-of-the-art systems for privacy. We consider methods of randomization, k-anonymization, and distributed privacy-preserving data mining. We also investigate cases in which the amount of data mining works to want to be sanitized for privacy-preservation purposes. We examine the computational and scientific goals related among privacy-preservation over high dimensional data sets.

We conducted a survey of the general areas of privacy preserving data mining and the underlying algorithms. We measured a variety of data compression methods such as randomization and k-anonymity based methods.

We examined methods for distributed privacy-preserving mining and the methods for handling horizontally and vertically distribution data. We examined the issue of reducing the effectiveness of data mining and data control works such as association rule mining, classification, and query processing. We examined any fundamental requirements of the problem of privacy preservation of the behavior of extended amounts of public data and training knowledge.

Finally, we analyzed a number of different application fields for that privacy-preserving data mining methods are useful.
2.2. Protected k-Nearest Neighbor Query on Encrypted Data in Outsourced Conditions

B.K. Samanthula proposed that the past decade, inquiry processing on relational data has been analyzed widely, and several technical and functional solutions to query processing have been offered under various scenarios. By the modern popularity of cloud computing, users presently have the possibility to outsource their data as well as the data control tasks to the cloud.

However, due to the addition of several privacy concerns, raw data (e.g., medical records) want to be encrypted since outsourcing to the cloud. In addition, query processing responsibilities should be controlled by the cloud; otherwise, where would be no point to outsourcing the data onto the first place. To prepare queries on encrypted data onto the cloud always decrypting the data is a highly challenging task.

We concentrate on explaining the k-nearest neighbor (kNN) query problem of encrypted database outsourced to a cloud: a user results from an encrypted query document to the cloud, and the cloud returns the k closest documents to the user. We first grant a fundamental design and confirm that such a simple solution is not secure. To implement better security, we propose a secure kNN protocol that protects the confidentiality of the data, user’s input query, and data access patterns. Also, we empirically investigate the performance of ours protocol through different experiments. These issues indicate that our secure protocol is highly efficient on the user end, and this lightweight design enables a user to use any mobile device to operate the kNN query.

2.3. k-Nearest Neighbor Classifier on Semantically Secure Encrypted Relational Data

In data mining uses classification is one of the generally used tasks. Earlier several solutions have been presented to the classification problem following different security models. This paper proposes a secure k-NN classifier on encrypted data in the cloud.

A novel privacy-preserving k-NN classification protocol, indicated by PPkNN is introduced. This is accomplished applying secure multiplication (SM), secure squared Euclidean distance (SSED), secure bit-decomposition (SBD), secure minimum and Secure Bit-OR (SBOR) protocols. PPkNN protocol protection the confidentiality of the data, user’s input queries, and protects the data access models.

3. EXISITING SYSTEM

We try to propose a Privacy preserving k-NN classifier algorithm. Here, the data is semantically secure and encrypted as well. Once, the information is encrypted and uploaded to the cloud the owner of the database does not include in any computations. So no data is revealed to the database owner. Thus, the content of database D and the query „q“ is not published to a cloud. Finally, the „Class Label: Cq“ is shown only to the approved user and no one else. In addition, later assigning his encrypted query report of the cloud, the user performs not include in any calculations.

Hence, data enters models are additional protected. These are the ways that protect the confidentiality and integrity of the data or information of the user.

Privacy preserving k-NN (D*, q) → C q privacy preserving K-NN is for privacy preserving K-NN protocol the protocol secures data confidentiality, user’s input query’s confidentiality, and further preserves the access models of data.
3.1. Mathematical Model

- Give S is every complete System that consists of
  - \( S = \{ Q, \text{privacy preserving K-NN}, D', \text{SRK-NN}, \text{secure Computation Majority class k,} \} \)
- Where denotes set of query registered by user
- \( D' = \text{Encrypted Data set.} \)
- PPKNN = process as privacy-preserving k-NN.
- SRKNN = Secure Retrieval of k-Nearest Neighbors.
- SCMCK = Secure Computation of Majority Class.

3.2. Algorithm

Here we present a collection of general sub-protocols that will continue utilized in building ours introduced k-NN protocol. All of the following protocols are viewed under two-client semi-honest connection. In particular, we investigate the performance from two semi-honest customers P1 and P2 such that the Palliser's cipher key SK is known only to P2 while it is public.

3.2.1. Secure Minimum Algorithm

At P1
1. Choose a random function of ‘F’
4. Generate public key using ‘F’
5. Aggregate lower properties
6. Send encrypted vectors to P2

At P2
7. Receive encrypted vectors from P1
8. Decrypt the encrypted vectors
9. P2 computes a new encrypted vector \( M^1 \)
10. At P1
11. Receive encrypted vectors from P2
12. Compute the inverse permutation of \( M^1 \)
13. P1 performs homomorphic operations

4. PROPOSED SYSTEM

4.1. PPkNN-algorithm

1. User encrypts his query q attribute wise, sends encrypted query about C1
2. C1 and C2:
   (a) C1 receives encrypted query and applies SSED protocol
   (b) C2 calculates the encryption applying SM protocol
3. (a) C1 and C2:
   C1 and C2 iteratively measures encryption compared
To k-nearest neighborhood of q
(b) C1:
At the end of the last iteration, only C1 knows the encryption key, sends to C2.
(c) C2:
Decrpts encrypted vectors and send to C1
(d) C1:
C1 Receives from C2 and performs inverse permutations
(e) Applying SBOR protocol C1 updates distance vectors with help of C2

4. Applying SCMC protocol C1 and C2 simultaneously computes KNN of q.

Explanation
The proposed PPkNN protocol essentially consists of the developing two stages:
Stage 1: Secure Retrieval of k-Nearest Neighbors (SRKNN):
• In this stage, User originally sends his query q (in encrypted form) to C1.
• Following this, C1 and C2 involve in a set of sub-protocols to securely retrieve (in encrypted form) the class.
• Labels communicating to the k-nearest neighbors of the input query q.
• By the end of this step, encrypted class labels of k-nearest neighbors are known only to C1.
Stage 2: Secure Computation of Majority Class (SCMCk):
• C1 and C2 simultaneously measure the class label by a majority voting amongst the k-nearest neighbors of q.
• At the end of this step, the unique User knows the class label corresponding to his input query record q.

In this system done by using ElGamal cryptosystem The ElGamal encryption system is an asymmetric key encryption algorithm for public-key cryptography which is based on the Diffie–Hellman key exchange. It was developed by Taher ElGamal in 1984. ElGamal is randomized, the trivial attack does not work, which is one of its advantages over RSA. Also, ElGamal uses a smaller key length when compared to RSA.

5. RESULTS
The results are shown in Fig.5 (a): Before encryption Fig.5 (b): After encryption Fig.5(c): K-means clustering

FIGURES

![Figure 1.1 k-NN classification](image-url)
Improvisation of k-NN Classifier on Semantically Secure Encrypted Relational Data

Figure 3.1 Protocol of PPkNN Classification

Figure 5 (a) Before encryption

Figure 5 (b) After encryption
6. CONCLUSION

This paper examines different current methods used for the privacy-preserving data mining (PPDM) and query approach on encrypted data. To secure client privacy, several privacy-preserving classification methods have been proposed over the previous decade. The popular methods are not applicable to outsourced database conditions anywhere the data continues 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. This protocol protects the confidentiality of the data, user’s input query, and protects the data access patterns. The future investigation can concentrate on further efficient solutions to the Secure MINn problem since the creation of the PPkNN protocol depends on the performance of the SMINn. Also, this paper can be increased towards the different classification algorithms.

REFERENCES

Improvisation of k-NN Classifier on Semantically Secure Encrypted Relational Data


