OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

REVIEW PAPER ON OVER SEMANTICALLY SAFE ENCRYPTED RELATIONAL DATA BY USING K-NEAREST NEIGHBOR CLASSIFICATION

DIPEEKA KANHIRAM RATHOD[1], ASST.PROF. V. S. KARWANDE[2]

ME CSE, Department of Computer Science & Engineering, EESGOI , India[1]
HOD, Assistant Professor, Department of Computer Science and Engineering, EESGOI, India. [2]

Abstract*: Data mining has a wide range of uses in a variety of industries, including banking, medicine, scientific research, and government agencies. One of the most common tasks in data mining applications is classification. Many theoretical and functional solutions to the classification problem have been suggested under various security models over the last decade as a result of the growth of various privacy issues. With the recent rise in popularity of cloud computing, users can now outsource their results, in encrypted form, as well as data mining tasks to the cloud. Current privacy-preserving classification methods aren't applicable since the data in the cloud is encrypted. The classification issue over encrypted data is the subject of this paper. We propose a stable k-NN classifier for encrypted data in the cloud, in particular. The proposed protocol safeguards data confidentiality, preserves the privacy of a user's query, and conceals data access patterns. To our knowledge, this is the first time a stable k-NN classifier has been developed over encrypted data using the semi-honest model. We also use a real-world dataset to empirically test the efficiency of our proposed protocol with various parameter settings. The K-nearest neighbor (KNN) classification method is widely used in data mining techniques. It is widely used in a variety of fields due to its ease of implementation, clarity of theory, and excellent classification efficiency. When training samples are distributed unevenly or the sample number of each class is very different, however, KNN can increase classification error rate. As a result, this paper adopts an improved KNN classification algorithm and applies it to object-oriented classification of high-resolution remote sensing images, building on the concept of clipping-KNN. Image artefacts are first collected as sample points by image segmentation. Second, the original KNN, clipping-KNN, and improved KNN are all implemented and used to classify the sample points. Finally, the effects of the classification are contrasted. The improved KNN algorithm can achieve higher precision in the classification of high-resolution remote sensing images in the same training and testing sets, according to the experiment.

Keywords: *K-Nearest Neighbor(KNN);KNN classification; object-oriented; segmentation; Data mining; privacy-preserving data mining (PPDM) ; stable multi-party computation (SMC) ; Machine Learning (ML).*

----------∴∴∴----------

I INTRODUCTION

The cloud computing paradigm has recently revolutionized how businesses manage their data, especially in terms of data storage, access, and processing. Many companies are seriously considering cloud computing as an evolving computing model because of its cost-efficiency, versatility, and offloading of administrative overhead. Organizations often delegate their computing operations to the cloud in addition to their data. Despite the many benefits that the cloud provides, businesses are unable to take advantage of such benefits due to privacy and security concerns. When data is extremely sensitive, it must be encrypted before being sent to the cloud. However, when data is encrypted, regardless of the underlying encryption scheme, conducting certain data mining tasks without first decrypting the data becomes extremely difficult. Other privacy issues exist, as shown by the following example.

When a user's record is part of a data mining operation, data mining over encrypted data (as described by DMED) on the

cloud must also secure the user's record. Furthermore, even though the data is encrypted, cloud can extract useful and confidential information about the individual data products by analyzing data access patterns. As a result, the DMED issue on the cloud has three privacy/security requirements:

(1) confidentiality of encrypted data,

(2) confidentiality of a user's query record, and

(3) hiding data access patterns.

The DMED problem cannot be solved by existing work on privacy-preserving data mining (PPDM) (either perturbation or stable multi-party computation (SMC) based approaches). Data perturbation techniques cannot be used to encrypt highly sensitive data since perturbed data lacks semantic protection. In addition, the skewed data does not yield very reliable data mining results. Data is distributed and not encrypted at each participating party in a secure multi-party computation-based approach. Furthermore, several intermediate computations are carried out using unencrypted data.

Assume an insurance firm outsourced to the cloud its encrypted customer database and related data mining activities. When a company agent needs to figure out the risk level of a potential new client, the agent may use a classification system. First, the agent must create a data record q for the customer, which contains the customer's personal information, such as credit score, age, marital status, and so on. Then you can send this record to the cloud, which will compute the class label for q. However, since q contains confidential information, it should be encrypted before being sent to the cloud to protect the customer's privacy.

"k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data" is a paper published in the journal Semantically Secure Encrypted Relational Data. The DMED problem cannot be solved by existing work on privacy-preserving data mining (PPDM) (either perturbation or stable multi-party computation (SMC) based approaches). Data perturbation techniques cannot be used to encrypt highly sensitive data since perturbed data lacks semantic protection. In addition, the skewed data does not yield very reliable data mining results. Data is distributed and not encrypted at each participating party in a secure multi-party computation based approach. Furthermore, several intermediate computations are carried out using unencrypted data.

The core idea behind kNN is that a test document is similar to a real document. planned to be in the same training class as the training documents that are found in the immediate vicinity examination. The majority class of a document's k closest relatives is assigned to it. next door neighbours It's crucial to choose a value for k because it affects the outcome. It has an effect on the classification's accuracy. It's preferable.to hold k at an odd number so that there are no ties if there are any. There are only two grades. Another choice is to adjust the value of k in such a way that the held-out part yields the best results of the training equipment.

If some classes have a large number of training documents, There's a fair chance that these will outperform the rest. Documents can be chosen from the k closest neighbours and the The paper will be immediately graded as an examination.Instead of the real class it belongs to, the plurality class is used. This limitation is solved by the proposed algorithm.

by first determining the size of the smallest class and then determining the size of the largest class.

Choosing the k closest neighbours :

One of the most common approaches in Artificial Intelligence is Machine Learning (ML). Machine Learning has been an important part of our lives over the last decade. It's used in tasks as basic as handwriting recognition and as complex as self-driving vehicles. It is also anticipated that the more mechanical repetitive tasks will be obsolete in a few decades. With the growing availability of data, there is reason to believe that Machine Learning will become an even more important component of technological advancement. Financial services, delivery, marketing and sales, and health care are only a few of the sectors where machine learning is having a significant effect. However, we will address Machine Learning's implementation and application in trading in this article.

K-Nearest Neighbors (KNN) is a basic machine learning algorithm for regression and classification problems. KNN algorithms take data and use similarity measures to identify new data points (e.g. distance function). A plurality vote of its neighbours is used to classify it. The information is allocated to the class with the most neighbours. As the number of nearest neighbours increases, so does the value of k, and so does the accuracy.

II LITERATURE SURVEY FOR SENTIMENT ANALYSIS

2.1. K-NEAREST NEIGHBOR (KNN) ALGORITHM FOR MACHINE LEARNING

The K-Nearest Neighbour algorithm is based on the Supervised Learning methodology and is one of the most basic Machine Learning algorithms. the K-NN algorithm assumes that the new case/data and existing cases are identical and places the new case in the category that is most similar to the existing categories. The K-NN algorithm stores all available data and classifies a new data point based on its

similarity to the existing data. This means that new data can be quickly categorised into a well-defined category using the K-NN algorithm. The K-NN algorithm can be used for both regression and classification, but it is most commonly used for classification problems. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data.

It's also known as a lazy learner algorithm because it doesn't learn from the training set right away; instead, it stores the dataset and performs an operation on it when it comes time to identify it. During the training process, the KNN algorithm simply stores the dataset, and when it receives new data, it classifies it into a group that is very close to the new data. Consider the following scenario: We have a picture of a creature that looks like a cat or a dog, but we don't know if it's a cat or a dog. We can use the KNN algorithm for this identification since it is based on a similarity measure. Our KNN model will look for similarities between the new data set and the cats and dogs images, and categorise it as either a cat or a dog based on the most similar features.

2.2. KNN CLASSIFICATION

Assume there are two categories, Category A and Category B, and we have a new data point x1. Which of these categories would this data point fall into A K-NN algorithm is needed to solve this type of problem. We can easily classify the type or class of a dataset with the aid of K-NN. Consider the diagram below:

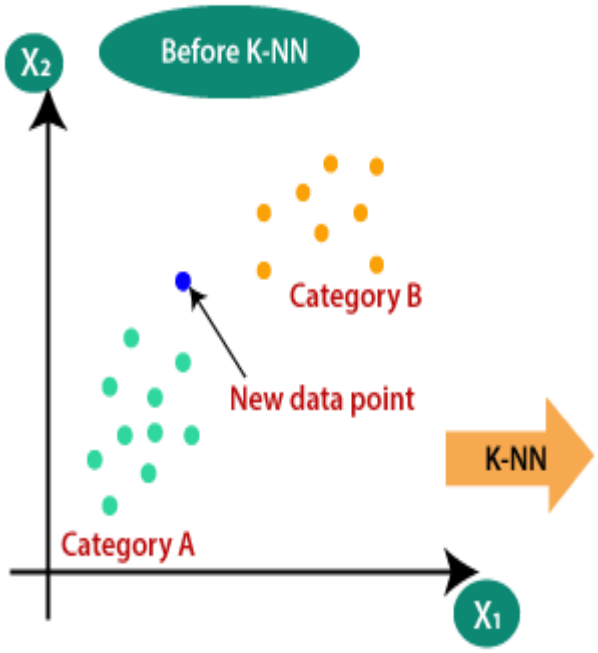

Figure No 2.1: K-NN Classification for New Data Point.

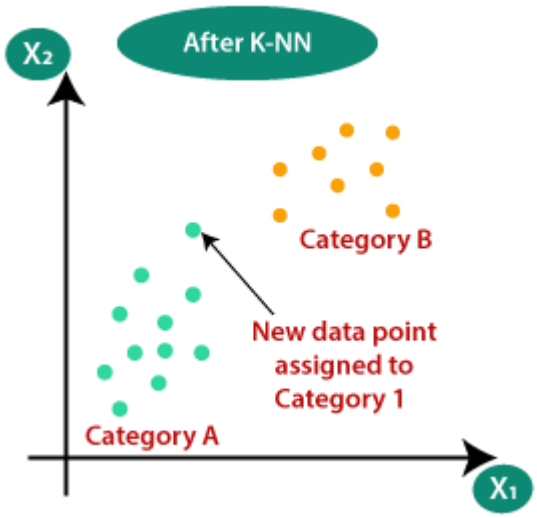

Figure No 2.2: Neighbours for New Data Point.

The following algorithm can be used to illustrate how K-NN works:

Stage 1: Settle on the number of neighbours (K).

Stage 2: Evaluate the Euclidean distance between K neighbours.

Step 3: Using the measured Euclidean distance, find the K closest neighbours.

Step 4: Count the number of data points in each group among these k neighbours.

Step 5: Assign the new data points to the group with the greatest number of neighbours.

Step 6: We've completed our model.

Let's assume we have a new data point that needs to be placed in the right group. Consider the following illustration:

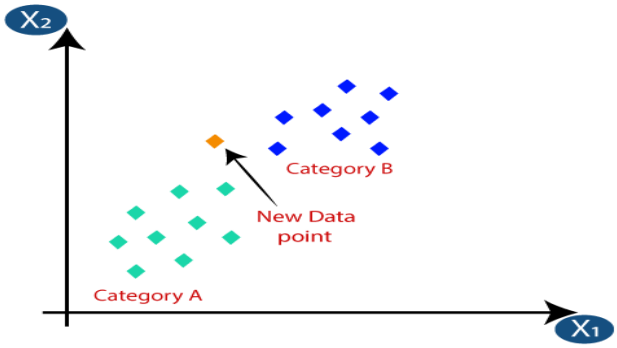

Figure No 2.3: Find New Data Point.

First, we'll decide on the number of neighbours, so we'll go with k=5.The Euclidean distance between the data points will then be calculated. The Euclidean distance is the distance between two points that we learned about in geometry class. It can be determined using the following formula:

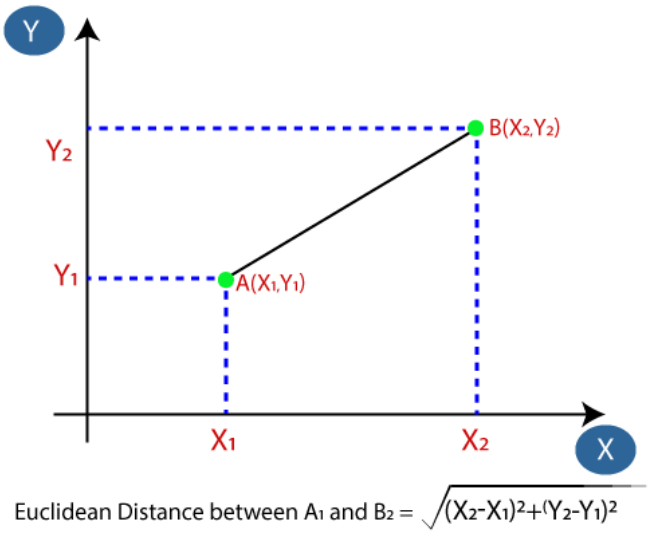

Figure No 2.4: Implementation for Euclidean Distance between two points

We found the closest neighbours by measuring the Euclidean distance, which yielded three closest neighbours in category A and two closest neighbours in category B. Consider the following illustration:

Figure No 2.5: Closest neighbours for two Category.

As can be shown, the three closest neighbours are all from category A, so this new data point must also be from that category.The following are some things to keep in mind when choosing the value of K in the K-NN algorithm:

There is no one-size-fits-all approach for deciding the best value for "K," so we'll have to play with a range of choices to find the best one. The value of K that is most commonly used is 5. K=1 or K=2 is a very low value for K that can be noisy and cause outlier effects in the model. Large values for K are desirable, but they can cause problems.

III. SYSTEMS ARCHITECTURE

Data owner register and login then browse and encrypt files and encrypt keyword and send the data to the Data server view the User details, attacker's details and unblock user. End User register and login then Search for files based on contents keyword, request for secret key, find the file based on SSED and K-Nearest neighbor search algorithm,

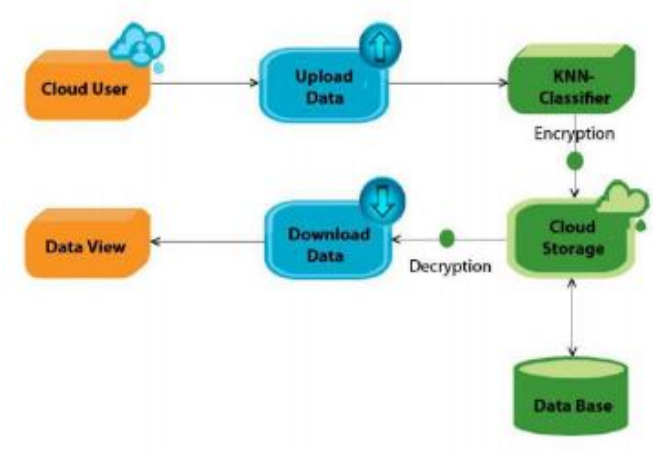

Figure No 3.1: Proposed System Architecture

IV CONCLUSION

Various privacy-preserving classification techniques have been proposed over the last decade to preserve user privacy. Current strategies are ineffective in outsourced database environments where data is stored on a third-party server in encrypted form. Over encrypted data in the cloud, this paper proposed a novel privacy-preserving k-NN classification protocol. Our protocol preserves the data's confidentiality, as well as the user's input question and data access patterns. We also checked our protocol's efficiency with various parameter settings. We expect to explore alternative and more effective solutions to the SMINn problem in our future work since improving the efficiency of SMINn is a significant first step in improving the performance of our PPkNN protocol. We'll also look at and build on our research into 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 Proc. 7th Int. Conf. Risk Security Internet Syst., 2012, pp. 1–9.

[3] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: Practical access pattern privacy and correctness on untrusted storage," in Proc. 15th ACM Conf. Comput. Commun. Security, 2008, pp. 139–148.

[4] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Proc. 17th Int. Conf. Theory Appl. Cryptographic Techn., 1999, pp. 223–238.

[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 Proc. 41st Annu. ACM Sympos. Theory Comput., 2009, pp. 169– 178.

[7] C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in Proc. 30th Annu. Int. Conf. Theory Appl. Cryptographic Techn.: Adv. Cryptol., 2011, pp. 129–148.

[8] A. Shamir, "How to share a secret," Commun. ACM, vol. 22, pp. 612–613, 1979.

[9] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy preserving computations," in Proc. 13th Eur. Symp. Res. Comput. Security: Comput. Security, 2008, pp. 192–206.

[10] R. Agrawal and R. Srikant, "Privacy preserving data mining," ACM Sigmod Rec., vol. 29, pp. 439–450, 2000.31.

[11] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Proc. 20th Annu. Int. Cryptol. Conf. Adv. Cryptol., 2000, pp. 36–54.

[12] P. Zhang, Y. Tong, S. Tang, and D. Yang, "Privacy preserving Naive Bayes classification," in Proc. 1st Int. Conf. Adv. Data Mining Appl., 2005, pp. 744–752.

[13] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Inf. Syst., vol. 29, no. 4, pp. 343–364, 2004.

[14] R. J. Bayardo and R. Agrawal, "Data privacy through optimal kanonymization," in Proc. IEEE 21st Int. Conf. Data Eng., 2005, pp. 217–228.

[15] H. Hu, J. Xu, C. Ren, and B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism," in Proc. IEEE 27th Int. Conf. Data Eng., 2011, pp. 601–612.

[16] M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier," in Proc. 8th Eur. Conf. Principles Practice Knowl. Discovery Databases, 2004, pp. 279–290.

[17] S. Zhang, M. Zong, K. Sun, Y. Liu, and D. Cheng, "Efficient kNN algorithm based on graph sparse reconstruction," in Proc. ADMA, 2014, pp. 356–369.