SECURE AND EFFICIENT DATA CLUSTERING AND CLASSIFICATION USING XOR ALGORITHM

[1]S. Ranichandra, [2]S. Dinesh, [3]R. Jeeva, [4]S. Selvakumari

Department of computer Science, Dhanalakshmi Srinivasan College of Arts and Science for Women,,Perambalur , 621 212, Tamilnadu, , India.

Email: ranichandras2523@yahoo.com (**S. Ranichandra***) Corresponding author:* S Ranichandra

ABSTRACT

When it comes to real-time data mining, there are several applications across a wide range of industries including finance, communications and biotechnology as well as the government. One of the most important aspects of data mining is classification. Because of the rise in privacy concerns, several theoretical and probable solutions to the categorization issue have been presented under various sureness models. Because the data in the data mining is encrypted, existing privacy protecting categorization algorithms are not connected to one another. Because the data in the data mining is encrypted, all privacy-preserving categorization techniques now in use are irrelevant. Based on the dataset, this research focuses on tackling the classification issue. Using the data mining strategy, we offer a safe k-NN classifier over encrypted data. Privacy of data, user input queries, and patterns of data access are protected by the proposed k-NN protocol. First of its kind in the standard using XOR encryption method, our work develops a safe k-NN classifier over encrypted data. An additional level of security is provided by using a secure kNN protocol that secures data, user input queries, and data access patterns. In addition, we conduct trials to test the efficacy of our processes. According to our studies, a user may utilise any mobile device to make a kNN query using our secure protocol, which is highly efficient on the user's end.

Key words: kNN Classifier, security, graph pattern matching, encryption, privacy preserving, secure protocol.

I. INTRODUCTION

Data mining is a powerful new method for sifting through massive amounts of data to find valuable information. Pattern recognition and numerical and mathematical approaches may also be used to identify new connections, patterns, and trends by analysing large volumes of data. The KNN-classification of time series is an important area of machine learning because of the vast volume of time-series data that is used in real-world applications. When it comes to storing, accessing, and processing information, the cloud computing concept is revolutionising N businesses' methods. Many organisations are drawn to cloud computing because of its affordability, adaptability, and reduction in administrative burden. In most cases, organisations outsource both their data and their computations to the cloud. Despite the many benefits of the cloud, corporations are unable to take use of them because of privacy and security concerns in the cloud. Encryption should be performed before data is sent to the cloud if the data is very sensitive, While it may be possible to do data mining operations without ever decrypting the encrypted data, this is not always the case [1]-[5]. .

Additional privacy issues have been verified by the following example. As an example, consider the following: Let's say an insurance firm decided to outsource the encryption and data mining of its customer database to the cloud. In order to identify the risk level of a potential new client, a categorization approach may be used by a corporate representative. A data record q for the client must first be created, which contains information such as the customer's credit score, employment location, age and marital status. As a result, the cloud may then use this documentation to figure out how many students q. needs to take the class. However, since q contains sensitive information, q must be encrypted before being sent to the cloud in order to preserve the privacy of customers. It is clear from the model above that data mining on a cloud containing encrypted data necessitates safeguarding a user's record. Even if the data are encrypted, the cloud can still get relevant and responsive information about what is really being accessed by observing the data access patterns. Therefore, the three privacy/security criteria of the cloud encryption challenge are: (1) privacy of the encrypted data, (2) secrecy of a user's query record, and (3) thrashing data access patterns. Existing work on Privacy-Preserving Data Mining (either perturbation or secure multi-party computing based on come close to) cannot overcome the encryption challenge [6]-[10].. This approach cannot be used to protect highly responsive data because troubled data lacks semantic security. Also, the data mining findings generated by the troublesome data aren't particularly accurate. This shift toward SMC requires that data be shared unencrypted among the parties involved. Non-encrypted data is used in many transitional calculations.

Encryption as a means of ensuring data privacy in the cloud may result in another problem during query processing. In general, it is very difficult to process encrypted data without having to decode it. The challenge here is how the cloud can run searches over encrypted data while the data stored in the cloud is always encrypted. Range queries and other collective inquiries have been suggested in the literature for query processing over encrypted data. However, the k-nearest neighbour (kNN) question cannot be answered using these approaches since they are either ineffective or inapplicable.

Classification is one of the most often used data mining techniques, and each approach has its own advantages. A security-based XOR technique is used to encrypt data in a cloud computing environment and perform k-nearest neighbour categorization [11]-[13].

Clustering

Clustering in data mining is necessary to classify patterns that are not immediately apparent. Traditional clustering algorithms are complicated by the characteristics of huge data: In the actual world, data is made of both numerical and category elements. Data that is a combination of both numerical and category kinds does not do well in most clustering techniques, whereas data that is solely numerical does.

For most clustering approaches, it is necessary to repeat the process many times before the grouping becomes better. Data mining applications may become unusable if the process takes an excessive amount of time.

Single-machine clustering methods and multiple-machine clustering techniques are the two most common types of data clustering. Because of their increased scalability and quicker response times for end users, multiple machine clustering approaches have recently gained a lot of interest. According to Fig. 2, clustering strategies that use one or more machines cover a wide range of methodologies:

Clustering on a single computer

Techniques that are based on samples

Measurement reduction methods

Clustering of many machines

Clustering in parallel

Clustering with Map Reduce

CONNECTED WORKS

[6] Fanyu Bu, Zhikui Chen, Qingchen Zhang, and Xin Wang [6] propose a k-means method based on partial distance (PDK algorithm). Imperfect information may be clustered using the PDK-means improved k-means method, which uses partial distance [4]. First, determine the distance between each item and each cluster so that objects may be assigned to clusters based on their proximity; second, figure out the mean value of each cluster so that the cluster centres can be updated. The execution time required by PDK-means is much smaller than that of PDPCM, especially when the data set is large, demonstrating that PDK-means achieve better for clustering incomplete high-dimensional big data. The PDK-means method outperforms the PDPCM algorithm in the tests. In addition, grouping incomplete high-dimensional huge data requires far less time than previously thought possible..

We originally proposed an unique density-based clustering technique, termed DBCURE, which is strong to locate clusters with various densities and suited for parallelizing the process using MapReduce. [12] In addition to discovering clusters with varied densities, DBCURE may also be used in a Map Reduce structure to find clusters. The parallel method DBCURE-MR was then devised and the correctness of DBCURE-MR was shown using Map Reduce. We demonstrated that our DBCURE-MR quickly detects precise clusters with unstable densities and scales up well with the Map Reduce structure by providing experiment results with varied datasets.

CURE's agglomerative HBC technique is presented by S. Guha, R. Rastogi, and K. Shim [14]. As soon as a partition has been generated, it is treated as a partial cluster in this technique. The last step is to eliminate outliers from each partition's data before creating final clusters. All of the cluster medoid (centroids) nearest to each other in a cluster are mixed in each stage of the method (merged). Only representative objects (centroids) will be combined in this technique, which employs a single linkage mechanism to choose several centroids from each cluster. It is impossible to quantify the joint interconnectivity of points in two distinct clusters of information while using this approach, which is a disadvantage. Chameleon algorithm [15] is used to get around this problem.

HBC agglomerative techniques, like as "CHAMELEON," may be used to decide dynamic modelling, according to E.-H. Han, V. Kumar, and G. Karypis [15]. A graph partitioning approach is used in the first step of this algorithm, which separates (partitions) data items into sub-clusters, before merging the sub-clusters and creating a final cluster. This approach is used to discover cluster densities in 2D (two-dimensional) space, where the clusters have varying forms and sizes. The Chameleon approach employs a dynamic model to create any cluster shapes and arbitrary cluster densities. The technique is useful for applications that deal with enormous amounts of data. Because CHAMELEON is known for its low dimensional data space, this is its biggest drawback.

Second, Data Mining That Protects Your Privacy (PPDM)

By using Privacy Preserving Data Mining, information about data may be extracted without compromising the privacy of the data (PPDM). Privacy-preserving categorization methods have been suggested in the literature during the last several decades. Based on a distributed training dataset, the goal of privacy preserving classification is to build a classifier that can predict the class label of an input data record.

Classification is a critical part of many data mining applications, including health care, retail, college, and business, among others. Data mining on the cloud has recently attracted much notice. Using cloud computing, a user puts his or her data in the hands of a cloud service provider. However, users' opinion is that privacy is a major concern when responsive data is outsourced to the cloud. The quickest approach to ensure the security of outsourced data is to encrypt it before it is sent to a third party.

Due to the following reasons, conventional privacy preservation classification approaches are not suitable or relevant to PPkNN, which hosts data in the cloud in encrypted form.

The data in present systems is divided between at least two parties, but the data in our instance is encrypted and kept on a cloud. (iii) Existing approaches are inaccurate owing to the loss of information due to the creation of statistical noises in order to hide the sensitive characteristic. It is possible for the cloud to get sensitive and important information about users' data items by just analysing the database entry patterns. In this study, we do not feel that the k-nearest neighbour approach, in which the data is discrete between two parties, is safe.

SYSTEM ARCHITECTURE

The system architecture consists of three practical components in Figure 1, Client, Data Server, and Backend Server.

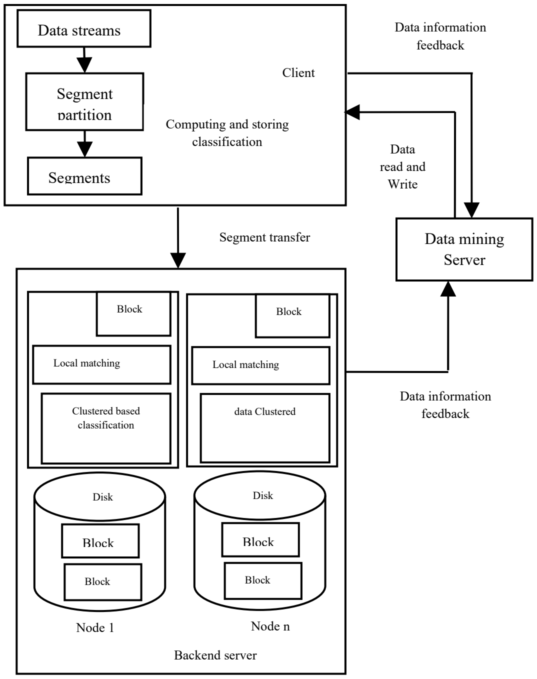

Fig 1 System Architecture

• The client gathers backup datasets and talks with the storage node and data server to replace information. In addition, the client does segmentation and data categorization, stores graph data via clustering, and distributes segments to storage nodes.

Data Server is responsible for storing and retrieving all graph index patterns matching files and segments.

In order to store the backup and to delete redundant data, backend servers are used. For parallel deduplication, the system calls for many storage nodes.

Implementation of our clustering-based categorization system. To distribute files to the nodes, a backup stream is split. Local graph index similarity is checked when a new data segment is received. The segment data is not saved if it is the same. The system refers to the block as "equivalent" if the similarity between it and the current block is strong. Similarity in graph index is checked when the data segment has a low degree of similarity. The approach assigns a data segment to a node if the data segments have a high degree of similarity. Home node stores and updates the clustered graph database server of other nodes if the segment data graph is not a high similarity is shown in figure 1

INTERNET CONNECTION

This module's primary purpose is to upload their data to the data server and store it there. The data owner encrypts the data file and then stores it on the server for security reasons. The owner of the data has the ability to decrypt and re-encrypt the file is shown in figure 2.

XOR IMPLEMENTATION

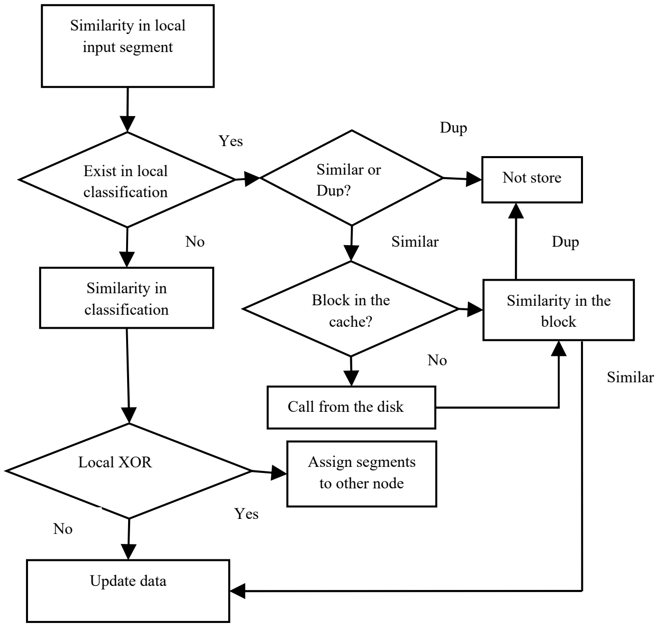

Figure 2 XOR implementation

EVALUATION

In order to measure our system's performance, we conduct memory and throughput tests. False positives are tolerated by the systems. The data sets are made up of files that have been backed up sequentially. Our approach is compared to two other situational algorithms, the Overlapping clustering algorithm and the slicing algorithm, for the elimination ratio. When testing memory for duplication eradication, the incremental backup is employed. As the number of nodes in the cluster grows, so does the experimental throughput for data categorization. Our system's performance will be evaluated via this experiment is shown in figure 3.

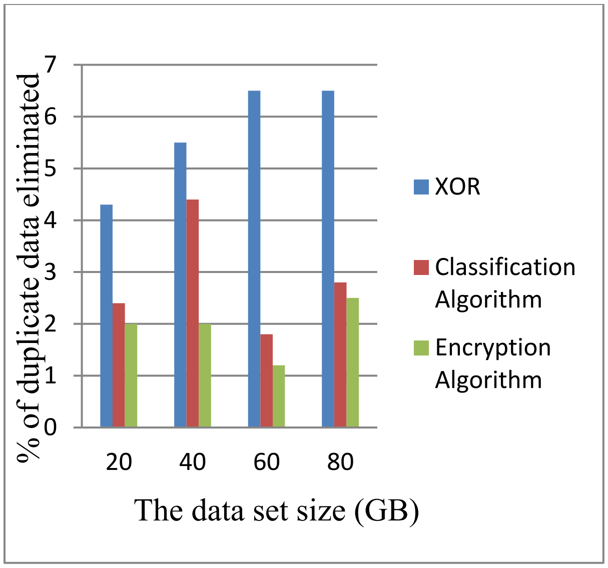

Figure 3 the percentage of performance

As the amount of the backup rises, we term our system XOR encryption; KNN with XOR improves roughly 60% to 68% of the data's speed. There is a 64.5 percent data rate ratio on average. Classification and encryption algorithms have average elimination rates of 28.38 percent and 23.23 percent, respectively. Our approach outperforms categorization and encryption algorithms when it comes to removing malware. Some data in the file is duplicated, and this is not exploited by overlapping clustering. A backup stream's inherent locality is used by an encryption technique. Because of this, if the data sets don't have a location, it'll lose its benefits and only find a small amount of redundant data. Our technique identifies more data duplications in a comparable data collection than any other. Pattern matching is used to discover comparable segments in the local graph pattern table, and a multi graph method is used to detect duplicate segments in the node.

CONCLUSION

As the amount of the backup rises, we term our system XOR encryption; KNN with XOR improves roughly 60% to 68% of the data's speed. There is a 64.5 percent data rate ratio on average. Classification and encryption algorithms have average elimination rates of 28.38 percent and 23.23 percent, respectively. Our approach outperforms categorization and encryption algorithms when it comes to removing malware. Some data in the file is duplicated, and this is not exploited by overlapping clustering. A backup stream's inherent locality is used by an encryption technique. Because of this, if the data sets don't have a location, it'll lose its benefits and only find a small amount of redundant data. Our technique identifies more data duplications in a comparable data collection than any other. Pattern matching is used to discover comparable segments in the local graph pattern table, and a multi graph method is used to detect duplicate segments in the node.

REFERENCES

[1] C. C. Aggarwal and P. S. Yu. A general survey of privacy-preserving data mining models and algorithms. *Privacy-preserving data mining*, pages 11–52, 2008.

[2] Y. Aumann and Y. Lindell. Security against covert adversaries: Efficient protocols for realistic adversaries. *Journal of Cryptology*, 23(2):281–343, Apr. 2010.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *ACM Sigmod Record*, volume 29, pages 439–450. ACM, 2000

[4] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Order preserving encryption for numeric data. In *ACM SIGMOD*, pages 563–574, 2004.26

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

[6] 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.

[7] Btissam Zerhari, Ayoub Ait Lahcen, Salma Mouline1, "Big Data Clustering: Algorithms and Challenges", CONFERENCE PAPER • MAY 2015F. Chung, Spectral Graph Theory. Providence, RI, USA: American Mathematical Society, 1997.

[8] S. Suthaharan, M. Alzahrani, ``Labelled data collection for anomaly detection in wireless sensor networks," in Proc. 6th Int. Conf. Intell. Sensors, Sensor Netw. Inform. Process. (ISSNIP), Dec. 2010, pp. 269_274.

[9] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, Sept. 1999.

[10] A. Katal, M. Wazid and R.H. Goudar, "Big data: Issues, challenges, tools and goodpractices," Contemporary Computing (IC3), 2013 Sixth International Conference on,IEEE, 2013.

[11] R. Xu and D. Wunsch, "Survey of clustering algorithms.," IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, vol. 16, no. 3, pp. 645-78, May. 2005.

[12] S.Kannadhasan and R.Nagarajan, Development of an H-Shaped Antenna with FR4 for 1-10GHz Wireless Communications, Textile Research Journal, DOI: 10.1177/00405175211003167 journals.sagepub.com/home/trj, March 21, 2021, Volume 91, Issue 15-16, August 2021, Sage Publishing

[13] S.Kannadhasan and R,Nagarajan, Performance Improvement of H-Shaped Antenna With Zener Diode for Textile Applications, The Journal of the Textile Institute, Taylor & Francis Group, DOI: 10.1080/00405000.2021.1944523