International Journal of Computing and Mathematics
Vol. 1, Issue. 2, June 2024, pp. 1~7

K-Means Clustering-based Privacy Preserving in Cloud Computing

Sri Wahyudi[1], Elviza Diana[2]
[1]STIKIP Rokania, Riau, Indonesia.
[2]Universitas Prof. Dr. Hazairin, SH, Bengkulu, Indonesia.
Email: sriwahyudi@rokania.ac.id[1], elvizaunihaz@gmail.com[2]

Article Info

Article history:

Received Apr 7, 2024
Revised Apr 29, 2024
Accepted May 28, 2024

Keywords:

Privacy-Preserving
k-means Clustering
Cloud Computing
Data Storage

ABSTRACT

A common data processing technique is clustering, which aims to divide information into related classes. Protecting database privacy is especially important when data is collected from various sensors. Scholars are working to limit the disclosure of personal data related to cloud computing because of significant elements that affect compliance with cloud information security. A k-means clustering approach is proposed in this research to safeguard privacy. To preserve user privacy in the online storage context, the clustering process does not reveal personal information or leak the variation matrix. Our security k-means primarily consist of a confidentiality cluster process computation. Two privacy-preserving techniques for clustering media computing are proposed in this research. Java is implementing our calculation method. Our version of the confidentiality method of clustering has been thoroughly tested on huge data sets. Results from experiments and theories confirm the security and accuracy of our process.

Corresponding Author:

Sri Wahyudi,
STIKIP Rokania, Riau, Indonesia.
sriwahyudi@rokania.ac.id.

1. INTRODUCTION

Collective data mining that protects privacy allows data mining methods to be jointly calculated without requiring those who are involved to reveal their specific data items to one another. The majority of the privacy-preserving protocols in the literature convert current data mining methods into privacy-preserving procedures. The resulting techniques may frequently leak fresh data [1]. Data mining techniques that protect privacy, such as those used by Lindell and Pinkas [3] and Agarwal and Srikant [2], force businesses to participate in collecting data without difficulty.The need that each party expose specific data pieces. Since the specifications of strategies-protected circuit-evaluation protocol are useless, several protected unique communications for specialist data mining algorithms were established.

Big data can now be stored and computed more efficiently thanks to recent advancements, but safeguarding consolidated data from various sources is still crucial [4]. Secure a coalition of parties computing has historically made confidentiality machine learning a high-level research subject. This allows various things to train distinct versions on their infinite memory while sharing everything except the data. Performance [5]. Due to the rapid development of desktop computers and wireless social media platforms, enormous amounts of data are constantly being produced from various mobile devices. and servers, posing a significant problem to corporations' computational capabilities [6]. The Cloud computing model is the, which is used to overcome this challenge. To provide economic costs, as more businesses store data in cloud servers; their efficient processing capacity is efficient for managing massive data quantities. Business analytics can indeed enable users and identify important information from a lot of information in companies and studies. Implementations. The study of this data makes it possible to forecast future developments and avenues for growth [7]. Clustering One of the primary research techniques in data analysis is to divide data

elements into many clusters, in this kind of way that the similarities between particles in a cluster, the similarity around each cluster is significant, meanwhile minimal.

A well-studied computational issue is clustering. To minimize an objective function, the objective is to aggregate the relevant ones into clusters in a standard normal distribution. The issue with the objective error- sum-of-squares is defined as the sum of the squares of the distance between nodes to their closest k clusters in the dataset. In clustering data processing, A massive portion of user-based privacy information is evaluated, such as geographic areas, data on carbon emissions, and spatial and temporal sensing data. The cybersecurity of this result of experiences regarding internet platforms' security.Essentially, a public cloud is used to store confidential data, which raises privacy concerns. of the applicant would also be uncovered if the provider of cloud storage is fraudulent. They integrate their tools to understand their specific distances if several users overlap. With each other and then measure the cluster centers by distance. Severe implications will arise if the privacy and cluster centers of users are exposed. As such, to secure the shared it can develop methodologies to preserve the privacy of users and cloud services. Such technology enables the cloud server to harvest social network information by delivering any user's privacy data while blocking all preventing users from simultaneously collecting data concerning another account or social media site.

Including participating users, the proposed privacy-preserving infrastructure focuses on safeguarding just one country's privacy from an online social network cluster that might expose intermediate data to possible cybercrime hackers or attackers unable to evade collusion attempts. A new approach called cloud computing makes use of its clustering characteristics to quickly identify processing needs and break them down into more manageable sub-requests. Performed and then deployed after installation by separate infrastructures. Cloud infrastructure is also a business-related material data analytics within the next five years should provide cloud service users with large-scale information system facilities. This service is known for its exceptional reliability, strong increase, cost reduction, and service on demand. It is widely agreed that Dropbox software features are the starting point of intrusion detection. In general, cloud infrastructure is based on the latest technology setting, which includes virtual machine protection, hardware security, and other network security concerns; these issues also occur in cloud computing security.

The query of k-clustering entails the division of information into k groups to limit the ESS. The notion of ESS can be used by Lloyd's (k-means) algorithm [1] regarding Ward's approach for multilevel agglomeration and k-clustering. In reality, It was found that Ward's algorithm functions fine, is quite slow (O(n3)), and doesn't expand well to large datasets. A variety of data mining algorithms have recently been described for inputs that are too big to fit completely in the main memory. In cloud storage security, an important technique for preserving order to get benefits is the processing of data before it is circulated. The infrastructure of that same spatial domain network retains effective measured data and internet connectivity by strict cryptographic guidelines. However, homomorphic encryption's computational price is high. This report argues a clustering algorithm towards privacy-preserving k-means to address these issues. About the suggested grouping algorithm, the k-means algorithm achieves adequate precision. Our strategy can avoid collusion attacks since all but one member is collaborating with the cloud server. The security accuracy of our method was confirmed by testing results as well as privacy studies.

2. RELATED WORKS

Personal space problems have been extensively studied in computational libraries. A very popular field of study has lately been privacy-preserving data mining. In this field, the initial emphasis was on the creation of decision trees from distributed data sets. There is also a large body of privacy research that maintains association principles for mining. This will concentrate on current work on privacy- preserving clustering.

Depending on inaccurate computation estimation and cryptographic techniques, Jha et al [8] suggested an algorithm for clustering privacy-preserving k-means. While this technique can be used in multi-party contexts, it also exposes the clustering centers to possible assaults on privacy. Bunn and Ostrovsky [9] have implemented A reciprocal grouping using a k-means approach that relies on cryptographic techniques to maintain the integrity of each statistic segment in which intermediate findings of the group and dataset distribution were never reported by the protocol. In particular, to randomly identify the original covariance matrix, a stable protocol was generated. As such, if the agreement grows to include multiple parties clustering with k-means, additional protection and privacy threats can be added.

For instance, there is no chance of plot assaults against the deal where more than half of the respondents partake in collusion. A Segmentation with a two-party k-means model was developed by Rao et al [10] to address this which could be done by the functional security of the technique of cryptographic techniques. Liu et al

[11] Suggested a secure approach for outsourced k mean grouping. Process to measure trap data; therefore, only one participant is bound to the privacy of this network. Using the asymmetric key encryption method in the k-means algorithm, Liu et al [11] extended to determine the sum Update the amount of data from sampling points in the cluster between each cluster and the clustering point. The multiplicative spatial domain encryption method is regarded in the literature review, as the multiplication method was used to determine the time using the formula. This solution ensures that by analyzing the k clusters, The geographical data is only discovered by the online supplier. Viewing the formula; respondents cannot access the intermediate knowledge of the students.

The multiplicative homomorphic encryption algorithm [12] had been suggested by existing literature; multiplicative homomorphism was shown to have some effect to support the Attack by Rivest Shamier-Adleman (RSA). A modern approach to sustaining the defense of the k- means clustering computing model based on the RSA multiplicative decryption has been suggested by some other studies [13]; this technique requires the RSA public-key cryptosystem and the key cryptographic protocol to preserve the data security of each user. The method of clustering is first calculated in the local center by each researcher via the k-means clustering algorithm and the effects are then encrypted. The local clustering outcomes are then retrieved from the data warehouse, and the latest study of cloud data analysis is terminated.

Zhang et al [14] offered a vectorized range to calculate the current average speed estimated by the interactive data set using encryption algorithms to access DPC cloud data. Adapting the algorithm to a multi-partner, however, the setting is unfeasible. The work of [15] recently unveiled a practical K-means clustering framework that might be easily transferred to cloud servers to protect privacy. They researched Map Reduce's stable incorporation into their system, making their system highly suited for the world of cloud computing. This work, however, exposes the clustering centers of the intermediate closet to the server. In the outsourcing context or differential privacy framework, several recent works concentrate on clustering. Few recent works suggest the protection of privacy by K-means clustering with complete assurances of privacy. The [16] method was performed with uniformly separated information only.

The distributed clustering of K-means [17] is based on the cryptographic protocol of Shamir, which incorporates two servers that do not collide with their system. The interpretation of the measurement metric in this study is therefore unclear. The techniques in [18] are not scalable for large scales and rely heavily on homomorphic encryption. Amounts of data. Commonly, hierarchical clustering protecting anonymity has been formally analyzed in [19]. The algorithm for hierarchical clustering is well known to have an O(n2 complexity where n represents the total amount of points of data (log(n)). K-means nowadays, which has an O(n) complexity and is greedy, is the most widely used clustering algorithm, but it has a drawback that we will address below Section. Thus, in this work, concentrate on the K- means algorithm's privacy-preserving approach.

Zhang et al. [20] suggested a high-order potentialistic c-means algorithm based on the BGV cryptosystem for big data in cloud computing. However, because of its low performance, their system is not realistic. Almutairi et al.

[21] made it successful and developed a privacy-saving k- means clustering mechanism focused on asymmetric cryptography, but it managed to secure plaintext information in the configuration of clustering cores. For this purpose, Yuan and Tian [22] also suggested a form of privacy-preserving clustering using a modern lightweight cryptosystem focused on error-learning difficulties. Using multidimensional data ciphertexts, their system will maximize the number of ciphertexts and compare the distance. This system is, however, not completely outsourced. In 1967, James MacQueen first used the conventional k-means [23].

Basu et al. [24] suggested presenting a pairwise restricted clustering structure and an efficient method of k-means for effects on organizational insightful pairwise constraints to enhance the efficiency of clustering. Researchers have recently proposed using particle swarm optimization (PSO) to boost the efficiency of the data cluster to improve k-means data clustering. The design and deployment of k-means clustering, also on large data sets, is quite simplistic. In a variety of topics, including consumer preference, computer vision, spatial analysis, physics, and agriculture, it has been used successfully.

3. PROPOSED K-MEANS CLUSTERING

One of the most predominant algorithms through unsupervised clustering is k-means, which could instantaneously partition a list of objects based on a certain similarity metric into k disjoint subsets. Typically, as clusters, can assign K disjoints subsets and require the Euclidean distance to determine the validity of the objects. The closer the Euclidean distance between two objects, in other words, the closer the distance between two objects, the greater the chance of clustering them into the same cluster, the more identical they are. There are many algorithms for clustering They possess unique advantages and downsides. The most widely employed algorithm in terms of constitutional analysis is K-means, which is inefficient and self-centered. In terms of computation [5]. Several steps are included in the K-means algorithm:

- Allocate every data point to the cluster with the closest centroid by calculating the amount of distance between all of them and every centroid.
- Update the centroid values by calculating the average of the point attribute values that are part of the cluster.

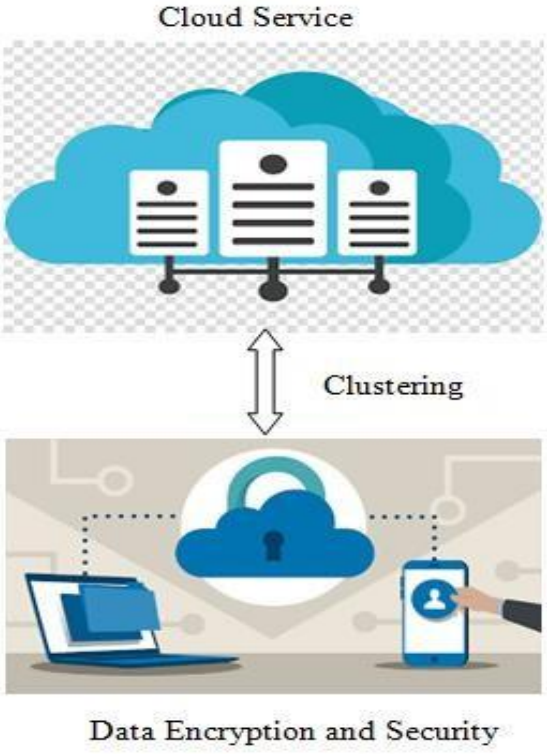

Figure 1. Clustering Framework

Our method functions in a classic "split, conquer, and combine" manner [1]. Using this methodology, the database can be divided into two equal portions. Then, k nearest neighbors can be extracted from each sector iteratively, and these two thousand clusters can be combined to create the final k clusters. With each recursive call, one can, nevertheless, identify a different separate tack, create 2k labels for each class, and then combine all 4k facilities into 2k centers. Ultimately, it came back from the recursion tree's top level to create the k final fields from the 2k clusters, that can be combined using the same technique. Following both recursive calls, the primary action of Combining 4,000 satellites into 2,000 fields is the result of To achieve this, we repeatedly choose and swap out the appropriate combination of hybridization Ci and Cj clusters in the cluster with Ci u Cj.

To create a structure with k- central components, a data owner trains their data using traditional k-means.

Points [25]. k clusters, for example. The data owner encrypts all k clusters employing the Ekey1() cryptography feature underneath a defined key1 before outsourcing to prevent the server in the cloud from learning the trained model. In contrast, the encryption outsourced model requires an owner of data with D-dimension information $x \in RD$ to attain a sample size for x. For secrecy, the data owner encrypts Ekey2(x) under key2 to conceal the true value of x from the website. Even though key1 and key2 are both invertible in our analysis, key1 is the inverted matrix, therefore they are uniformly represented as SK symmetric keys in the following data encryption, the data user provides the appropriate digital signature to the cloud server, which groups the encrypted information into the appropriate group according to a privacy-preserving distance comparison. Therefore, those who claim that comparing confidentiality distance is the most effective method to Our approach could solve our encrypted k-Means issue efficiently by employing scalar-product-preserving cryptography.

In our article, The basic methodology used to establish K- indicates that the safety feature is scalar product-preserving cryptography. which had been presented in [26]. Except maybe for applying a lower color

of random numbers against important " models, our architecture does not alter encryption construction. Hence, as long as encryption is structurally stable, the adversary cannot retrieve the plaintext from its ciphertext as well as our system is stable. It neglects the security properties of the original encryption to prevent replication and instead includes a privacy review under various threat models for our suggested K-means.

4. EXPERIMENTAL RESULTS

To provide In these segments, a clustering technique that protects anonymity using our suggested approach and to clarify the test case findings. It also reviews the effectiveness of the suggested state-of-the-art privacy-preserving proposed approaches. Users search it on a single server with 2x 36-core Intel Xeon 2.30GHz CPUs and 256GB RAMM to find optimization of our model. While there are various cores, amongst each group it does it's solely using one thread for calculation. We use a single connection to run both companies., but to simulate a wireless router, we use the Linux tc command: 0.02ms round trip latency LAN world, 10 Gbps bandwidth network. It notes With linear cost analysis, it is hard to guess running times on the WAN as the combined running time is simply the sum of computer time and data transmission time. The previous job, in relation, only performed experimental numbers in the LAN environment. Thus, in all the tests below, it will concentrate on the LAN environment.

This paper presents in this section, the simulation data regarding our clustered strategy that protects privacy our tests demonstrate accessibility and scalability by using a global collection of artificial data sets. As opposed to earlier research, we also compare our scheme with the actual dataset. The accuracy is the number of individuals correctly clustered in the assessment package. In this section, using our suggested methodology and Compare the results of the designed system that uses a technique for grouping K-means in plain text. To improve understanding, we utilize the 2D information from off [27] and S1 [28], which contain the precise point cluster labels or centers. The plain-text technique is frequently tested, as is our framework of free speech, and views the centroids of the collected sets. Except for the update process, all the functions used in our method essentially turn the fractional part of the current centroid cluster into integers just like the preparatory algorithms used for K-means clustering for plain text. Conclude that the standardization strategy outlined in [29] is applied. The experimental findings indicate that relative to the original function, normalization has a marginal effect on model specificity. K-means clustering on decimal numbers equivalent to plain text, our system of normalization achieves the same precision.

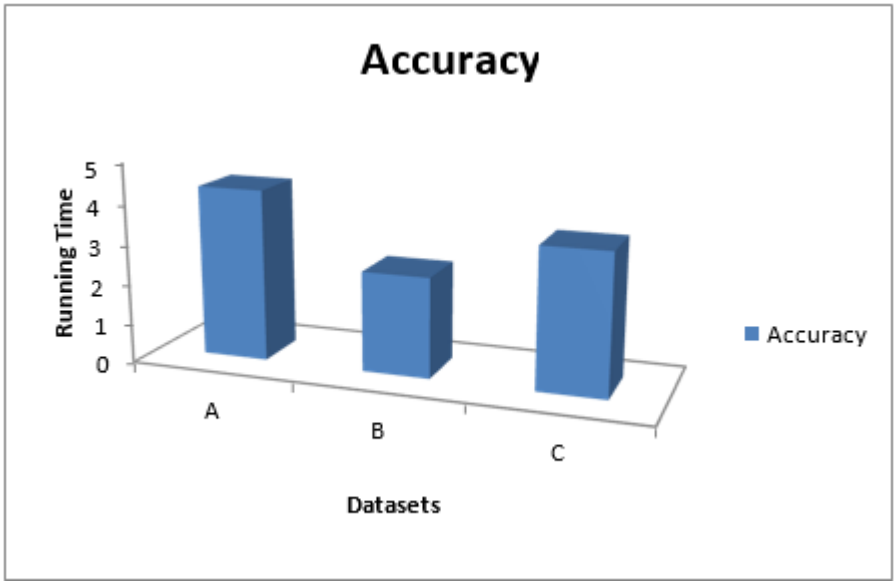

Figure 2. Performance of Privacy-Preserving K-Means Clustering Algorithm

Moreover, such errors still occur in the Algorithm K- means itself. A noted deficiency of the K-means algorithm is that its accuracy exceeds consistency. At the initialization phase, a random set of data points, resulting in separate clusters as the algorithm, is carried in a local optimum and therefore does not contribute to the optimum global at the activation phase. As such, you should still operate the algorithm with various vector clustering, and then pick the test result that generates the lowest number of square distances. Therefore, the plaintext k-mean algorithm on S1 is also contrasted with our privacy-preserving model.

Some of you might know which centroid fits them the best based on the ground truth of database S1. The Euclidean distance between each centroid and all the true centroids is examined, and each centroid is mapped to the actual ground reality centroid whose smallest distance from all grounding truth centroids is. The consistency of our privacy-preserving method is equal to that of the plaintext technique in all tests. The following is a definition of this: the demise of our protocol's legitimacy is due largely to the garble-circuit-based-division process as listed above, compared to the plaintext algorithm. Therefore, in our tests, the loss of precision does not happen. However, if this error is not very small, we would like to assume that The quality of our model may be obtained using digits used in the method of division. Also, if we kept more digits during the truncation stage, the accuracy of our system would have increased, but it requires extensive computation/communication quantities. There is a tradeoff that would be between the application of the division's computing time and accuracy.

5. CONCLUSION

The issue of confidentiality leakage in clustering regulations is addressed in this research by proposing an efficient privacy-preserving k-means method that securely determines the optimal grouping hubs for every participant sans exposing the participants to any clustering information. Furthermore, the subject is not aware of the privacy particulars of other subjects in the same cluster. None of the other participants' private data is disclosed, despite a systematic attack detection, revealed even though colluding participants exist. No participant may achieve other participants' or cluster centers' private information. In comparison, cloud service providers measure cluster centers without understanding the private details of the participants. Experimental findings on the 3 data sets showed that, relative to the hierarchical clustering techniques and basic k-means algorithm, Good time efficiency and clustering impacts have been seen in the proposed algorithm. It will understand and adopt the safeguards of privacy of other clustering algorithms to solving scenarios in future studies.

REFERENCES

[1] Jagannathan, G., Pillaipakkamnatt, K., & Wright, R. N. (2006, April). A new privacy-preserving distributed k-clustering algorithm. In Proceedings of the 2006 SIAM international conference on data mining (pp. 494-498). Society for Industrial and Applied Mathematics.

[2] Agrawal, R., & Srikant, R. (2000, May). Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 439-450).

[3] Lindell, Y., & Pinkas, B. (2002). Privacy preserving data mining. Journal of cryptology, 15(3).

[4] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

[5] Mohassel, P., Rosulek, M., & Trieu, N. (2020). Practical privacy-preserving k-means clustering. Proceedings on Privacy Enhancing Technologies, 2020(4), 414-433.

[6] Sun, L., Ci, S., Liu, X., Zheng, X., Yu, Q., &

[7] Luo, Y. (2020). A privacy-preserving density peak clustering algorithm in cloud computing. Concurrency and Computation: Practice and Experience, 32(11), e5641.

[8] Bhatia, T., Verma, A. K., & Sharma, G. (2020). Towards a secure incremental proxy re-encryption for e-healthcare data sharing in mobile cloud computing. Concurrency and Computation: Practice and Experience, 32(5), e5520.

[9] Jha, S., Kruger, L., & McDaniel, P. (2005, September). Privacy preserving clustering. In European symposium on research in computer security (pp. 397- 417). Springer, Berlin, Heidelberg.

[10] Bunn, P., & Ostrovsky, R. (2007, October). Secure two-party k-means clustering. In Proceedings of the 14th ACM conference on Computer and communications security (pp. 486-497).

[11] Rao, F. Y., Samanthula, B. K., Bertino, E., Yi, X., & Liu, D. (2015, October). Privacy-preserving and outsourced multi-user k-means clustering. In 2015 IEEE Conference on Collaboration and Internet Computing (CIC) (pp. 80-89).IEEE.

[12] Liu, D., Bertino, E., & Yi, X. (2014, June). Privacy of outsourced k-means clustering. In Proceedings of the 9th ACM symposium on Information, computer and communications security (pp. 123-134).

[13] Rivest, R. L., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2), 120- 126.

[14] Yuan W, Ren X. Research on privacy preserving clustering method for horizontal partitioned data. J Comput Technol Develop. 2015;5:115-117.

[15] Zhang, Q., Zhong, H., Yang, L. T., Chen, Z., & Bu, F. (2016). PPHOCFS: Privacy preserving high-order CFS algorithm on the cloud for clustering multimedia data. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 12(4s), 1- 15.

[16] Yuan, J., & Tian, Y. (2017). Practical privacy- preserving mapreduce based k-means clustering over large-scale dataset. IEEE Transactions on Cloud Computing.

[17] Gheid, Z., & Challal, Y. (2016, August). Efficient and privacy-preserving k-means clustering for big data mining. In 2016 IEEE Trustcom/BigDataSE/ISPA (pp. 791-798). IEEE.

[18] Dimitrakos, T., Moona, R., Patel, D., & McKnight, D. H. (Eds.). (2012). Trust Management VI: 6th IFIP WG 11.11 International Conference, IFIPTM 2012, Surat, India, May 21-25, 2012, Proceedings (Vol.374). Springer.

[19] Xing, K., Hu, C., Yu, J., Cheng, X., & Zhang, F. (2017). Mutual privacy preserving $ k $-means clustering in social participatory sensing. IEEE Transactions on Industrial Informatics, 13(4), 2066-2076.

[20] Meng, X., Papadopoulos, D., Oprea, A., & Triandopoulos, N. (1904). Privacy-Preserving Hierarchical Clustering: Formal Security and E icient Approximation.

[21] Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2017). PPHOPCM: Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing. IEEE Transactions on Big Data.

[22] Almutairi, N., Coenen, F., & Dures, K. (2017, August). K-means clustering using homomorphic encryption and an updatable distance matrix: secure third party data clustering with limited data owner interaction. In International Conference on Big Data Analytics and Knowledge Discovery (pp. 274-285). Springer, Cham.

[23] Yuan, J., & Tian, Y. (2017). Practical privacy- preserving mapreduce based k-means clustering over large-scale dataset. IEEE Transactions on Cloud Computing.

[24] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[25] Basu, S., Banerjee, A., & Mooney, R. J. (2004, April). Active semi-supervision for pairwise constrained clustering. In Proceedings of the 2004 SIAM international conference on data mining (pp. 333-344). Society for Industrial and Applied Mathematics.

[26] Yin, H., Zhang, J., Xiong, Y., Huang, X., & Deng, T. (2018). PPK-means: Achieving privacy- preserving clustering over encrypted multi-dimensional cloud data. Electronics, 7(11), 310.

[27] Wong, W. K., Cheung, D. W. L., Kao, B., & Mamoulis, N. (2009, June). Secure kNN computation on encrypted databases. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 139-152). https://github.com/deric/clustering-benchmark.

[28] Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. Applied Intelligence, 48(12), 4743-4759.

[29] Mohassel, P., & Zhang, Y. (2017, May). Secureml: A system for scalable privacy-preserving machine learning. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 19-38). IEEE.