International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Privacy Preserving Access Controlled Interactive Clustering as Service Over Hybrid Cloud

Amogh Pramod Kulkarni[1], Dr. Manjunath T. N.[2], Shwetha Shetty B.[3], [4]Chandrashekara Lingaiah Nagaratna

Submitted: 04/12/2023 **Revised**: 12/01/2024 **Accepted**: 27/01/2024

Abstract: Mining on large volume of data offloaded by enterprises to cloud has become an integral part in business strategy design. But the mining must be privacy preserving as leakage of data or mined information can create various security and privacy threats. Towards this end, privacy preserving data mining techniques can become critical. This work proposes a privacy preserving clustering model with support of incremental and adaptive clustering, fine grained access control and prevention from leakage of data and security parameters. The clustering is built on privacy preserved locality sensitive hashing technique on binary vector summarized data. The privacy of security parameter is ensured using generative adversarial network (GAN) deep learning model. Through experimental analysis, the proposed solution is found to reduce the computation cost for clustering by 40%, communication cost by 12% and able to provide better clustering accuracy with ARI of about 5% compared to existing works.

Keywords: *Clustering, Deep Learning, Generative Adversarial Network, Privacy preserving,*

1. Introduction

In era of data revolution, large volume of historical data collected by enterprises is a huge wealth for them. These data have hidden patterns about consumer and sales and valuable business insights can be mined from this data. These business insights are needed for effective business strategization. With emergence of cloud computing, many enterprises are shifting their storage and computations to data center as it provides various benefits like reduced cost, on demand scaling etc. Offloading storage and computation to third party datacenter, though brings many benefits, it also exposes the data to various privacy and security vulnerabilities. Private and sensitive data can be leaked directly or through inference. Most existing works on privacy preserving cluster transformed the data and executed clustering over the transformed data. These methods assume trusted cloud and data transformation operations were executed on cloud. In presence of curios public cloud, the data transformation overhead is on user end to prevent data leakage. Since most data transformation operations are based on homomorphic encryption and the cloud can infer inherent data characteristics even though the data is transformed. The communication overhead and clustering latency is higher for incremental adaptive clustering in existing works due to multiple interactions between the user and the cloud end. Inference attacks can also launch on these interactions to learn data characteristics. This work proposes a privacy preserving access controlled interactive clustering over the hybrid cloud. Recently hybrid cloud has been explored as a prospective solution for solving the security and privacy issues in public cloud in a scalable and computationally efficient manner. Hybrid cloud solves the security and privacy issues by distributing sensitive data to private cloud and in-sensitive data to public cloud. The solution proposed in this work is built on this hybrid cloud concept of data distribution based on data sensitivity. In the proposed solution the data is summarized using a novel data summarization algorithm and a locality sensitivity hash is constructed on it for privacy preserving interactive adaptive clustering. The original data is transformed using attribute-based encryption and stored in public cloud. The binary vector summarization is kept at private cloud and used for clustering. The security parameters used for data transformation are prevented from internal attacks using Generative Adversarial Networks (GAN). GAN transforms the security parameters to a mask image and stores in private cloud. Following are the novel contributions of work.

1. A novel interactive adaptive privacy preserved clustering algorithm on hybrid cloud framework based on locality sensitive hash of data. The algorithm is privacy preserved with facility for adapting the number of clusters and attributes for encryption.
2. The security parameters used for data transformation is secured using GAN. By this way, security parameters are prevented from leakage by internal attackers. GAN transforms the security parameters to mask image which is hard to decipher even if leaked.

[1] *Research scholar, Visvesvaraya Technological University, Belagavi, Asst. Prof., dept. of Artificial intelligence & machine learning, B.M.S. College of Engineering, Bengaluru.*
[2] *Professor, department of Information Science and engineering, B.M.S. Institute of Technology and Management, Bengaluru.*
[3] *Senior Business Data Analyst, Loblaw companies ltd, Canada.*
[4]*Senior Software Principal Engineer, Dell Technologies*
kulkarni84@gmail.com

Paper organization is as follows. Section II presents the existing privacy preserving clustering algorithms and review on their applicability to cloud computing environment. The proposed methodology of privacy preserving access controlled interactive clustering is presented in Section III. The results of the proposed solution and comparison with existing works are presented in Section IV. Finally, the section V provides the concluding remarks and scope of further research.

2. Related Work

A privacy preserving clustering scheme based on bipartite graph was proposed for big data by Zobaed et al [1]. The scheme is for unstructured data where cluster centroids are formed on topic diversity of texts. Data items are associated to cluster based on weighted bipartite graph association to cluster centroids. Privacy of data item is ensured by transforming to tokens. The token vocabulary space is huge for unstructured texts, due to which clustering complexity is high. Also the approach lacks support for interactive adaptive clustering. A fast version of clustering using K-means without iterations was proposed by Wei fu et al [2]. K folder cross validation is used to select the cluster centers. The clustering can be done over transformed data for privacy preservation. Though privacy preservation is possible in this approach, it does not support interactive adaptive clustering and fine-grained control over clustering. Dynamic and incremental clustering was proposed by Mary et al [3] by modifying the DBSCAN clustering algorithm. Though the approach addressed incremental clustering for handling streaming data, privacy and fine-grained control over attributes for clustering was not considered in this work. Double encryption-based privacy preserving clustering algorithm was proposed by Rong et al [4]. Arithmetic and equality operations were executed on transformed data. But the computation complexity is very high in this approach and it does not support interactive clustering. Privacy preserving K-means clustering for cloud data was proposed by Yuan et al [5]. Though the method is able to solve the problem of data leakage by curious cloud service providers, it has high data transformation overhead on the data owners. Data owner needs to compute the cluster centroids and pass it cloud, opening the communication channel for inference attacks. Homomorphic encryption (HE) was used in the privacy preserving K means clustering algorithm proposed by Rao et al [6]. Data owner uploads the HE encrypted data to cloud and clustering is done over encrypted data. But the method does not support interactive adaptive clustering. Also, curious cloud providers can learn data characteristics through inference. BCP double encryption with additive homomorphic property is used for privacy preserving clustering by Zou et al [7]. Double encryption secures the data from data inference attacks by curious cloud providers. The computation complexity is high and the method is not scalable for large volume of dataset. Also, the method is not suitable for interactive adaptive clustering. Differential privacy integrated K means clustering was proposed by Yu et al [8]. Data transformation is done by adding Laplacian noise to the done. The noise was added differentially based on the data sensitivity. Though privacy is ensured, the method does not support interactive adaptive clustering. Also it does provide access control at fine grained level. Optimized canopy algorithm (OCA) combined with K-means clustering algorithm is used for differential privacy by Shang et al [9]. OCA selects the initial centroids and privacy preserving K means builds on it for further clustering. But the scheme does not support interactive adaptive clustering. A privacy preserving clustering scheme over data transformed using BGV homomorphic encryption was proposed by Zhang et al [10]. Though privacy was ensured, the scheme does not support interactive adaptive clustering and differential privacy to the users. Privacy preserving clustering using randomized kernel matrix for data perturbation was realized by Lin et al [11]. Though privacy is ensured, the approach does not support interactive adaptive clustering. Privacy preserving K means algorithm for horizontally partitioned dataset administered by multiple parties was proposed by Gheid et al [12]. Each party does cluster on its data and send the cluster centers after additive transformation. From multiple cluster centers, new cluster centers are found. The scheme is not suitable for interactive adaptive clustering. Privacy preserving K means clustering algorithm on negative transformed dataset was proposed in works of Hu et al [13], Zhao et al [14] and Zhao et al [15].

In the negative transformed database, distance computation is realized as bit difference binary operation. The method is not suited for incremental adaptive clustering and does not support fine grained access control. A distributed privacy preserving K mean clustering algorithm was proposed by Brando et al [16]. Clustering was done at local sites and cluster centroids are encrypted using HE. From encrypted centroids, the global centroid is computed and shared to local sites. Though privacy of cluster centroid is preserved, interactive adaptive clustering is not supported by this work. HE was used for realizing privacy preserving clustering in works of Jiang et al [17] and Almutairi et al [18]. Data is transformed using HE and uploaded to cloud. Clustering is realized on the transformed data. But the method is not secure against curious cloud service provider. Also, the method does not support interactive adaptive clustering. Archana et al [38] explored various data masking methods for securing data. But data utilities like clustering cannot be executed as these methods are not order preserving. Ravi Kumar et al [39] proposed bit manipulation-based data masking with order preserving property but the method is not scalable and does not support access control.

From the survey, it could be seen that most of the privacy preserving clustering does not support interactive adaptive

clustering and fine-grained access control. Most of the data transformation schemes are based on HE and it is insecure against curious cloud service providers. Attackers can learn data characteristics from the HE data. In most HE based approaches computation overhead is higher on data owner end.

3. GAN Based Privacy Preserving Over Hybrid Cloud

Hybrid cloud is used in this work to address the multiple requirements of privacy preservation clustering, interactive/ adaptive clustering and access control and security of data/security parameters. The architecture of the solution is given in Figure 1. The proposed solution has three important functionalities- data transformation, privacy preserving interactive clustering and security parameter transformation. The data is transformed into a attribute value encoding scheme at private cloud. The transformed data is kept at public cloud. The keys and access control parameters are transformed using GAN and stored at private cloud to prevent from internal attacks. To facilitate privacy preserving interactive adaptive clustering, data is summarized to binary vector and clustered using locality sensitive hash (LSH) at private cloud. User interacts with private cloud, to cluster based on bucket transformation in an interactive adaptive manner. The details of each of the functionality is presented below.

A. Data transformation

The data transformation operation is handled at private cloud. The data transformation technique proposed in this work is designed to support incremental data. In addition to data transformation, a shortened representation of data for purpose of privacy preserving interactive adaptive clustering is generated. The shortened representation is generated using LSH. The transformed data is uploaded to public cloud. The shortened representation is kept at private cloud.

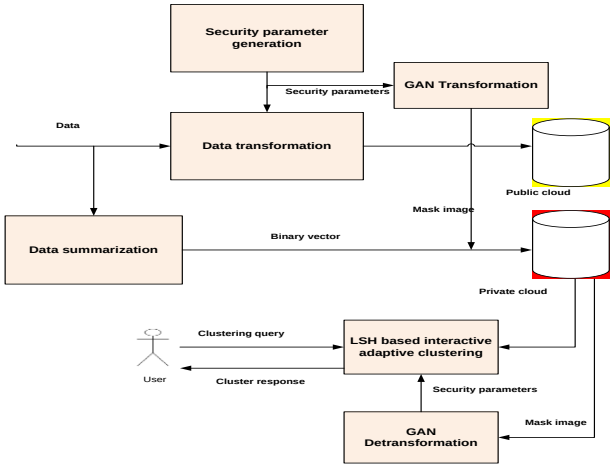

Fig. 1. Architecture of proposed privacy preserving interactive clustering

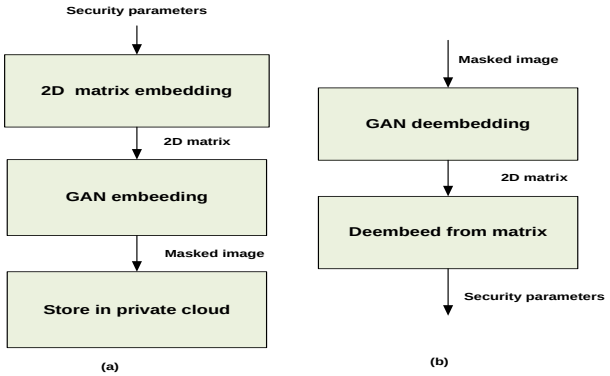

Fig. 2. (a) GAN transformation (b) GAN de-transformation

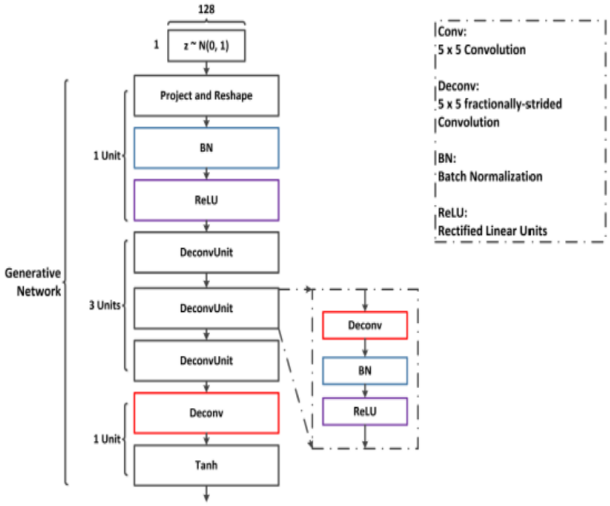

Fig. 3. Generative network

The data uploaded by data owner is in table format in which each column is an attribute. The attributes are marked as sensitive or in-sensitive by the data owner. The sensitive attributes are encrypted using attribute-based encryption (CP-ABE). The user access attributes over which the access control on data must be enforced is defined by the data owner. System administrator at private cloud generates a public key (PK) and master key (MK)as

$$PK = \mathbb{G}_0, g, h = g^{\beta}, f = g^{\frac{1}{\beta}}, e(g,g)^{\alpha} \quad (1)$$

$$MK = (\beta, g^{\alpha}) \quad (2)$$

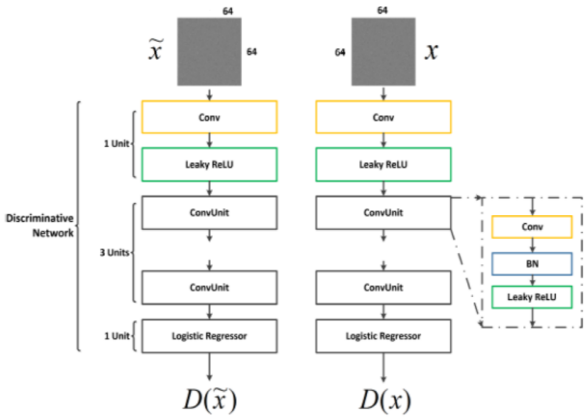

Fig. 4. Discriminator network

Where $\mathbb{G}_0$ is the bilinear group of prime order p with generator g. α, β are two random prime numbers. The sensitive attributes (M) are encrypted using PK and access tree (T) as

$$C = (\mathrm{T}, \bar{\mathrm{C}} = \mathrm{Me}(\mathrm{g}, \mathrm{g})^{\alpha s}) \quad (3)$$

Where s is the random prime number. The data table with sensitive attributes replaced by encrypted sensitive attributes is kept at public cloud. Differing from earlier works of using HE for encrypting the data, this work uses attribute-based encryption with non-additive properties to secure the data in public cloud from inference attack by curious cloud providers. For interactive adaptive clustering a shorted representation of data is created using LSH with a data summarization procedure.

Locality sensitivity hashing (LSH) [22] is technique for reducing the search time. It does this reduction by mapping the high dimensional space to a lower dimension space using group of hash functions. The advantage of LSH is that, it hashes the item repeatedly to a location in the address space, so that semantically similar items are mapped to closer location in the address space. The hash function construction process in LSH is given as

$\mathrm{g}: \mathrm{R}^{\mathrm{d}} \rightarrow \mathrm{U}$ such that for any two points x, y

if $||\mathrm{x} - \mathrm{y}|| \leq \mathrm{r}$, then $\Pr[\mathrm{g}(\mathrm{x}) = \mathrm{g}(\mathrm{y})]$ is high

if $||\mathrm{x} - \mathrm{y}|| > cr, then\ \Pr[\mathrm{g}(\mathrm{x}) = \mathrm{g}(\mathrm{y})]$ is small

This is achieved with family H of functions

$\mathrm{g}(\mathrm{x}) =< \mathrm{h}_1(\mathrm{x}), \mathrm{h}_2(\mathrm{x}), \ldots . \mathrm{h}_{\mathrm{k}}(\mathrm{x}) >$

For all data point xϵP, p is hashed to buckets $\mathrm{g}_1(\mathrm{x}), \mathrm{g}_2(\mathrm{x}), \ldots \mathrm{g}_{\mathrm{b}}(\mathrm{x})$

For an input query y, the points are retrieved from the buckets $\mathrm{g}_1(\mathrm{y}), \mathrm{g}_2(\mathrm{y}), \ldots$ until all points from b buckets are retrieved.

Every time when a batch of data arrives at private cloud, data summarization procedure is invoked.

The data summarization procedure is as follows. Say the data has N rows with d attributes. In each row, the attributes are first normalized as in Eq 2.

$$x_{ti} = \frac{F(x_i)}{max_i - min_i} \quad (4)$$

In the equation (4), the attribute value is represented as $F(x_i)$. This function converts the categorical value to numerical value by sorting the categorical value in some order and returning the index of the location as result. The value of x_{ti} is first rounded to 10 decimal and then converted to binary vector using the equation (5).

$$\mathrm{B}(x_{ti}) = \begin{cases} 1, x_{ti} * 10 \\ 0, elsewhere \end{cases} \quad (5)$$

A binary vector is thus created for every attribute value. The binary vectors are hashed using LSH into buckets. The binary vector and mapping of binary vectors to data items is kept at private cloud.

B. Interactive adaptive clustering

User interacts with public cloud for clustering. User provides the attribute list and the number of clusters as input. This input is processed at private cloud. The binary vectors are masked to zero in the regions where attributes are not requested for clustering. After masking LSH is done over binary vectors to result in K buckets. The buckets are either merged, split or kept same depending on the number of clusters (N) requested by the user. The bucket processing rules are given in below table

Table 1. Bucket processing rules

$N == K$	Buckets are kept same
$N < K$	The average binary vector is created for all items in the bucket by doing bitwise AND operation. Hamming distance is calculated between the average binary vector of each bucket and the bucket with lowest hamming distance are merged. The merging process is repeated till $N < K$
$N > K$	For each bucket maximal hamming distance between binary vector is found. The bucket with maximal hamming distance is split into two by clustering to two groups. This split process is repeated till $N > K$

Once the N buckets are found, the median binary vectors which is close to all binary vectors in the same bucket is found. The binary vector which is minimum distance to median $(minI)$ and maximum distance to median $(maxI)$ is found. For each bucket, the data items corresponding to the

$minI$ and $maxI$ are downloaded from the public cloud and sent to the user. The data items sensitive attributes are still encrypted. The user can decrypt the sensitive attributes as

$$M = C/\left(\frac{e\left(h^s, g^{\frac{\alpha+r}{\beta}}\right)}{e(g,g)^{rs}}\right) \quad (6)$$

From the decrypted $minI$ and $maxI$ data item, the minimum and maximum values for each of the attributes requested by the user are provided, so that user can get the characteristics of data distribution in each cluster.

C. Security parameter transformation

The security parameters of attribute-based encryption are kept in private cloud. These parameters can be compromised by internal attacks at private cloud. A privacy preserving scheme using GAN is proposed in this work for ensuring the security of data transformation of security parameters. GAN maps the data transformation parameters to a mask. This mask is then saved to private cloud. The advantage of GAN is that, mask created by GAN is very secure and difficult to decipher.

GAN consists of two networks – generative and discriminative. The configuration for the generative and discriminative networks used in this work is given in Figure 3 and Figure 4. The goal of generative network is to generate synthetic samples in such a way to deceive the discriminative network. The goal of discriminative network is to find if the sample is real or creation of generative network. With both generative and discriminative network competing, the generated synthetic samples by the generator are close to real.

The objective function GAN is given as

$$L_{GAN} = E_{\bar{x}\sim P_g}[D(\bar{x})] - E_{\bar{x}\sim P_r}[D(x)] + \lambda E_{\bar{x}\sim P_x}[(||\nabla_{\bar{x}} D(\bar{x})||_2 - 1)^2] \quad (7)$$

The distribution over V is given as P_r. The distribution over which generator produces data is given as P_g. The uniform samples over P_r and P_g is given as P_x.

The overall process for security parameter transformation using GAN is given in figure 2(a). The security parameters are embedded into a 2D matrix. This matrix is passed to GAN embedding to generate the masked image from the 2D matrix. This masked image is kept at private cloud instead of plain storage of key parameters. Thus, even if the masked image is leakage due to internal attacks, it becomes difficult for attacker to learn the secret parameters embedded in the image.

The process for security parameter retrieval is given in figure 2(b). The masked image is passed to GAN de-embedding to get the 2D matrix. From the 2D matrix, the security parameters are retrieved.

4. Results

Experimental evaluation of the proposed privacy preserving algorithm was done against four real world datasets of: Household electricity power consumption (HPC), Forest cover and KDD-99 downloaded from UCI machine learning repository.

The performance of the proposed solution is evaluated in terms of (i) purity, (ii)Adjusted random index (ARI), (iii) computation time and (iv) communication overhead.

Purity (P)is used to measure the clustering accuracy. It is calculated for K cluster as

$$P = \frac{\sum_{i=1}^{K} \frac{|c_i^d|}{|c_i|}}{K} \times 100 \quad (4)$$

In the equation (4), the cardinality of the dominant cluster is represented as $|c_i^d|$.

Clustering accuracy by comparing to ground truth is measured using two parameters of Adjusted Rand Index (ARI) [35] and Purity. ARI is measured by calculating the similarity between the clustering results and the ground truth as

$$ARI = \frac{RI - groundtruth(RI)}{(\max(RI) - groundtruth(RI)} \quad (5)$$

In the equation (5), the RI is given

$$RI = \frac{a+b}{\langle {n \atop 2} \rangle} \quad (6)$$

In the equation (5), the number of matching pairs between the clustering result and ground truth is represented as ‘a’ and the number of non-matching pairs between the clustering result and ground truth is represented as ‘b’ and n is the total number of elements. ARI value ranges from zero to one. Value towards one denotes best clustering and value towards zero represents poor clustering.

For comparing the performance of the proposed solution, PPCOM solution proposed in [4] and secure nearest neighbor clustering (SNN) proposed in [20] as used.

The comparison of purity values across the solutions is given in Table 2.

Table 2 Comparison of purity

Solutions	HPC	Forest cover	KDD-99	Average
Proposed	0.90	0.91	0.92	0.91
PPCOM	0.84	0.86	0.87	0.85
SNNC	0.82	0.86	0.84	0.84

On an average, the proposed solution has 6% higher purity compared to PPCOM and 7% higher purity compared to SNNC. Use of LSH has provided faster adaptivity to

incremental clustering requirements in proposed solution compared to PPCOM and SNNC. This has increased the accuracy of clustering in proposed solution and shown as higher Purity value.

ARI is measured for various datasets and the result is given in Table 3.

Solutions	HPC	Forest cover	KDD-99	Average
Proposed	0.89	0.90	0.89	0.89
PPCOM	0.83	0.85	0.84	0.84
SNNC	0.81	0.82	0.81	0.81

On an average, proposed solution has 5%, 8% ARI compared to PPCOM and SNNC respectively. ARI has increased in proposed solution due to cluster merging or splitting based on hamming distance variation of binary vector summarization.

The cluster computation time is measured for various cluster number for various datasets and the result is given in Figure 5-7.

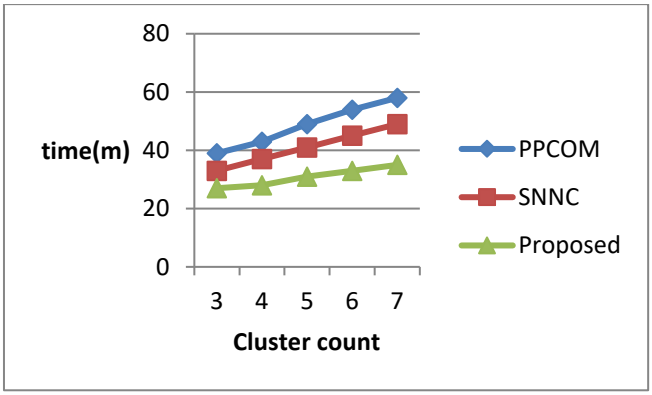

Fig. 5. Cluster computation time for HPC dataset

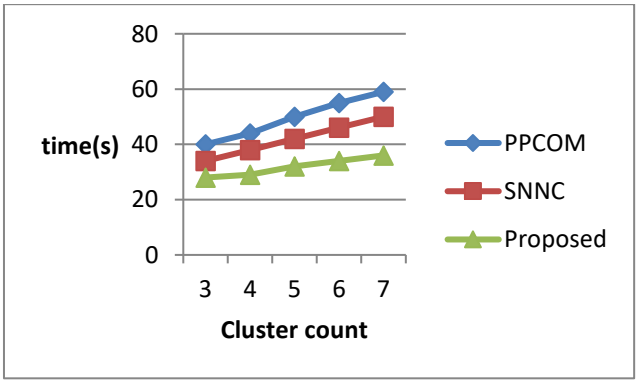

Fig. 6. Cluster computation time for Forest cover dataset

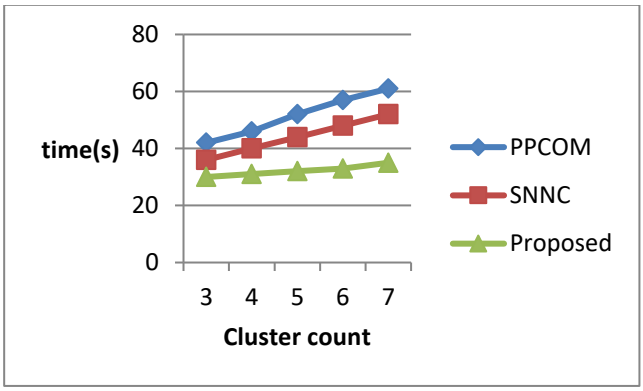

Fig. 7. Cluster computation time for KDD-99 dataset

The difference in time between cluster count 3 to 7 on average is 19s in PPCOM, 16s in SNNC and 8s in proposed solution. Clustering time increment is lower in proposed solution compared to PPCOM and SNNC. This is due to merging and splitting of buckets for clustering in proposed solution rather than iteration based clustering in PPCOM and SNNC.

The communication overhead for clustering is measured for various number of clusters for different datasets and the result is given in Figure 8-10.

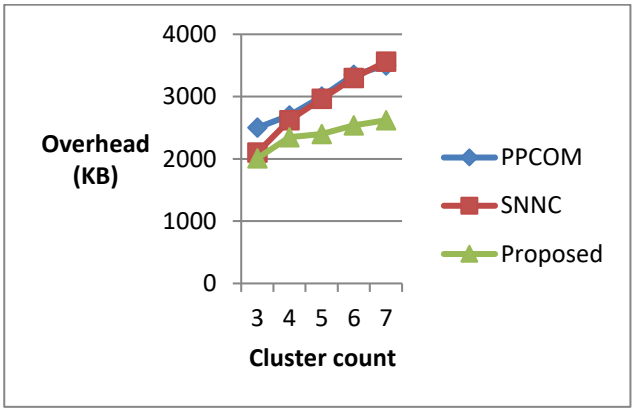

Fig. 8. Communication overhead for HPC dataset

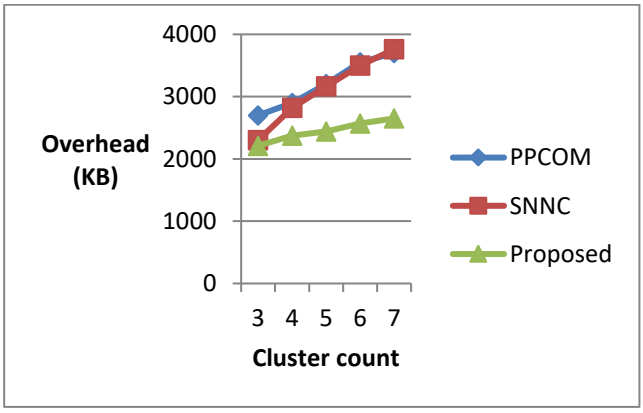

Fig. 9. Communication overhead for Forest cover dataset

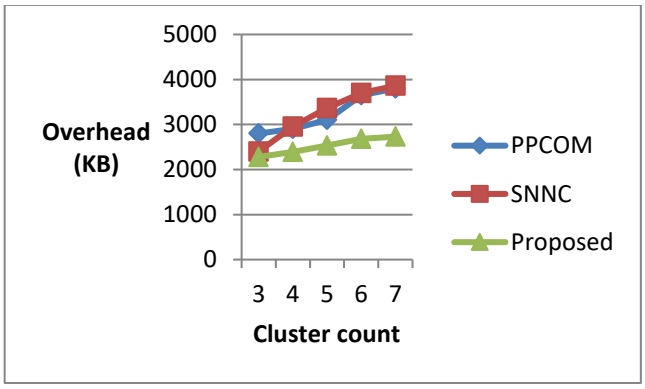

Fig. 10. Communication overhead for KDD-99 dataset

The difference in communication overhead between cluster 3 to 7 on average is 1000KB in PPCOM,1460KB in SNNC and 610KB in proposed solution. Communication overhead is reduced in proposed solution due to reduction is volume of data communicated between the user and the cloud service provider for interactive clustering.

The difficulty in predicting the original data from the transformed data is represented as security strength and it is calculated by measuring the difference between the original data and predicted original data from transformed data by the attacker over the period of time. The variance of difference between the predicted and original data is given as VoD and security strength is calculated in terms of VoD as

$$ss = \frac{\sum_{i=1}^{N} VOD_i}{N}$$

The value of ss is calculated for 5-hour duration for various guess by the attacker and plotted in a 1-hour scale as given in Figure 11.

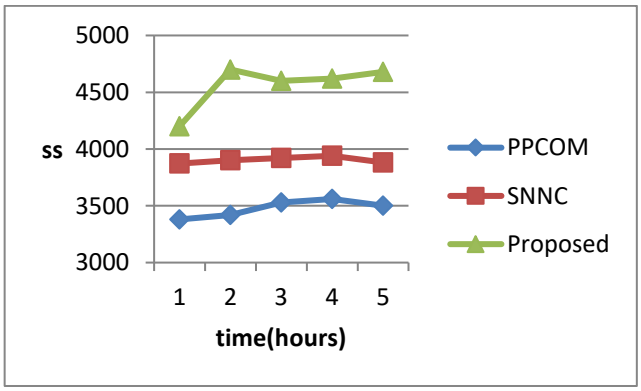

Fig. 11. VoD plot

The ss is higher in proposed solution compared to SNNC and PPCOM. This is due to use of attribute based encryption in proposed solution which is stronger compared HE based solution used in PPCOM and SSNC.

The VoD is measured for security parameter transformation using GAN based transformation in proposed solution and result is given in Figure 12. The VoD for security parameter transformation is very high indicating that it is very difficult to decipher the security parameters from the GAN generated masked image.

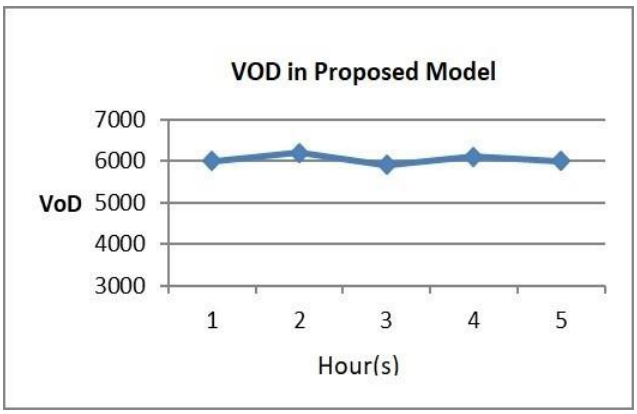

Fig. 12. VoD for security parameters

5. Conclusion

A privacy preserving access controlled interactive clustering solution based on hybrid cloud is proposed in this work. The solution transformed the data using attribute based encryption to provide fine grained access control to the users. For the data equivalent data summarization is done using a novel data summarization technique to facilitate interactive clustering using LSH. The security parameters for data transformation are prevented from internal attacks using GAN masking. The proposed solution had comparatively lower computation and communication cost for clustering. The computation complexity is lower in the proposed solution as it avoids complex cryptographic primitives and iterative process in clustering. By using minimal communication between private and public cloud, the communication overhead is also lower in the proposed solution. The proposed solution can be extended for hierarchical clustering model as part of future work.

Author contributions

Amogh Kulkarni: Conceptualization, Literature survey, Methodology, Data Curation, Writing-Original draft preparation**. Manjunath T N:** Validation, Field study. **Shwetha Shetty:** Data Curation, Visualization, Investigation, **Chandrashekara Nagaratna** Validation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Zobaed, S., Gottumukkala, R.N., & Salehi, M. (2020). Privacy-Preserving Clustering of Unstructured Big Data for Cloud-Based Enterprise Search Solutions. *ArXiv, abs/2005.11317*.

[2] Wei Fu and Patrick O Perry. Estimating the number of clusters using cross-validation. Journal of Computational and Graphical Statistics, pages 1–12,

January 2019.

[3] Angel Latha Mary and KR Shankar Kumar. A density based dynamic data clustering algorithm based on incremental dataset. Journal of Computer Science, 8(5):656–664, February 2012

[4] Hong Rong, Huimei Wang, Jian Liu, Jialu Hao, Ming Xian, "Privacy-Preserving -Means Clustering under Multiowner Setting in Distributed Cloud Environments", Security and Communication Networks, vol. 2017, Article ID 3910126, 19 pages, 2017

[5] J. Yuan and Y. Tian, "Practical Privacy-Preserving MapReduce Based K-Means Clustering Over Large-Scale Dataset" in IEEE Transactions on Cloud Computing, vol. 7, no. 02, pp. 568-579, 2019

[6] F.-Y. Rao, B. K. Samanthula, E. Bertino, X. Yi, and D. Liu, "Privacy-preserving and outsourced multi-user k-means clustering," in Proceedings of the 1st IEEE International Conference on Collaboration and Internet Computing, CIC 2015, pp. 80–89, October 2015.

[7] Ying Zou, Zhen Zhao, Sha Shi, Lei Wang, Yunfeng Peng, Yuan Ping, Baocang Wang, "Highly Secure Privacy-Preserving Outsourced k-Means Clustering under Multiple Keys in Cloud Computing", Security and Communication Networks, vol. 2020

[8] Q. Yu, Y. Luo, C. Chen, and X. Ding, "Outlier-eliminated k-means clustering algorithm based on differential privacy preservation," Applied Intelligence, vol. 45, no. 4, pp. 1179–1191, 2016.

[9] T. Shang, Z. Zhao, Z. Guan, and J. Liu, "A DP canopy k-means algorithm for privacy preservation of hadoop platform," in Proceedings of the CSS 2017, Lecture Notes in Computer Science, vol. 10581, pp. 189–198, Springer, Xi'an, China, October 2017

[10] H. Rong, H. Wang, J. Liu, J. Hao, and M. Xian, "Outsourced k-means clustering over encrypted data under multiple keys in spark framework," in Proceedings of the SecureComm 2017, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 238, pp. 67–87, Springer, Niagara Falls, ON, Canada, October 2017.

[11] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "Pphopcm: privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing," IEEE Transactions on Big Data, 2017.

[12] N. Almutairi, F. Coenen, and K. Dures, "K-means clustering using homomorphic encryption and an updatable distance matrix: secure third party data clustering with limited data owner interaction," in Proceedings of the DaWaK 2017, Lecture Notes in Computer Science, vol. 10440, pp. 274–285, Springer, Lyon, France, August 2017.

[13] K.-P. Lin, "Privacy-preserving kernel k-means clustering outsourcing with random transformation," Knowledge and Information Systems, vol. 49, no. 3, pp. 885–908, 2016

[14] Z. Gheid and Y. Challal, "Efficient and privacy-preserving k-means clustering for big data mining," in Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA, pp. 791–798, IEEE, Tianjin, China, August 2016.

[15] Hu Xiaoyi,Lu Liping,"Privacy-Preserving K-Means Clustering Upon Negative Databases",Springer International Publishing,2018

[16] D. Zhao *et al*., "A Fine-grained Privacy-preserving k-means Clustering Algorithm Upon Negative Databases," *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 1945-1951

[17] Dongdong Zhao, Xiaoyi Hu, Shengwu Xiong, Jing Tian,"k-means clustering and kNN classification based on negative databases",Applied Soft Computing,2021

[18] Brandão, André & Mendes, Ricardo & Vilela, Joao. (2021). Efficient Privacy Preserving Distributed K-Means for Non-IID Data. 10.1007/978-3-030-74251-5_35.

[19] Jiang, Z.L., Guo, N., Jin, Y., Lv, J., Wu, Y., Liu, Z., Fang, J., Yiu, S.M., Wang, X.: Efficient two-party privacy-preserving collaborative k-means clustering protocol supporting both storage and computation outsourcing. Information Sciences 518, 168–180 (2020)

[20] Almutairi, N., Coenen, F., & Dures, K. (2018). Third Party Data Clustering Over Encrypted Data Without Data Owner Participation: Introducing the Encrypted Distance Matrix. *DaWaK*.

[21] X. Xu and X. Zhao, "A Framework for Privacy-Aware Computing on Hybrid Clouds with Mixed-Sensitivity Data," *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, 2015

[22] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In Proceedings of the twentieth annual symposium on Computational geometry. 253–262.

[23] Mayank Bawa, Tyson Condie, and Prasanna Ganesan.

2005. LSH forest: self-tuning indexes for similarity search. In Proceedings of the 14th international conference on World Wide Web. 651–660

[24] Junhao Gan, Jianlin Feng, Qiong Fang, and Wilfred Ng. 2012. Locality-sensitive hashing scheme based on dynamic collision counting. In Proceedings of the 2012 ACM SIGMOD international conference on management of data. 541–552.

[25] Qiang Huang, Jianlin Feng, Yikai Zhang, Qiong Fang, and Wilfred Ng. 2015. Query-aware locality-sensitive hashing for approximate nearest neighbor search. Proceedings of the VLDB Endowment 9, 1 (2015), 1–12

[26] Wanqi Liu, Hanchen Wang, Ying Zhang, Wei Wang, and Lu Qin. 2019. I-LSH: I/O efficient c-approximate nearest neighbor search in highdimensional space. In 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, 1670–167

[27] Yifang Sun, Wei Wang, Jianbin Qin, Ying Zhang, and Xuemin Lin. 2014. SRS: solving c-approximate nearest neighbor queries in high dimensional euclidean space with a tiny index. Proceedings of the VLDB Endowment (2014).

[28] McConville, R., Cao, X., Liu, W., & Miller, P. (2016). Accelerating Large Scale Centroid-based Clustering with Locality Sensitive Hashing. In Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering (ICDE) (pp. 649-660). Institute of Electrical and Electronics Engineers (IEEE).

[29] C. Oprişa, M. Checicheş and A. Năndrean, "Locality-sensitive hashing optimizations for fast malware clustering," *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2014, pp. 97-104

[30] Jafari, Omid & Maurya, Preeti & Nagarkar, Parth & Islam, Khandker & Crushev, Chidambaram. (2021). A Survey on Locality Sensitive Hashing Algorithms and their Applications.

[31] Khader, Mariam & Al-Naymat, Ghazi. (2020). Density-based Algorithms for Big Data Clustering Using MapReduce Framework: A Comprehensive Study. ACM Computing Surveys. 53. 1-38. 10.1145/3403951.

[32] A. Andoni and P. Indyk, "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions," *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 2006, pp. 459-468

[33] M. Hahsler, M. Bolanos, and J. Forrest, stream: Infrastructure for Data Stream Mining, 2015, R package version 1.2-2

[34] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary? in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 1073–1080

[35] Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846–850

[36] L. Wan, W.K. Ng, X.H. Dang, P.S. Yu, K. Zhang, Density-based clustering of data streams at multiple resolutions, ACM Trans. Knowl. Discov. Data 3 (3) (2009) 49–50.

[37] Xu, Ji & Wang, Guoyin & Li, Tianrui & Deng, Weihui & Guanglei, Gou. (2017). Fat Node Leading Tree for Data Stream Clustering with Density Peaks. Knowledge-Based Systems. 120. 99-117. 10.1016/j.knosys.2016.12.025.

[38] Archana, R.A. & Hegadi, Ravindra & T N, Manjunath. (2017). A Big Data Security using Data Masking Methods. Indonesian Journal of Electrical Engineering and Computer Science. 7. 449-456. 10.11591/ijeecs.v7.i2. pp449-456.

[39] G K, Ravikumar & T N, Manjunath & Hegadi, Ravindra. (2011). Design of Data Masking Architecture and Analysis of Data Masking Techniques for Testing. International Journal of Engineering Science.