Data and Metadata. 2024; 3:287
doi: 10.56294/dm2024287

ORIGINAL

LDCML: A Novel AI-Driven Approach for Privacy-Preserving Anonymization of Quasi-Identifiers

LDCML: Un nuevo enfoque basado en IA para anonimización de cuasidentificadores para preservar la intimidad

Sreemoyee Biswas[1] ✉, Vrashti Nagar[1] ✉, Nilay Khare[1] ✉, Priyank Jain[2] ✉, Pragati Agrawal[1] ✉

[1]Computer Science, Maulana Azad National Institute of Technology. Bhopal, 462003, Madhya Pradesh, India.
[2]Computer Science, Indian Institute of Information Technology. Pune, 412109, Maharashtra, India.
These authors contributed equally to this work.

Cite as: Biswas S, Nagar V, Khare N, Jain P, Agrawal P. LDCML: a novel ai-driven approach for privacy-preserving anonymization of quasi-identifiers. Data and Metadata. 2024; 3:287. https://doi.org/10.56294/dm2024287

Submitted: 14-10-2023 **Revised:** 17-02-2024 **Accepted:** 16-05-2024 **Published:** 17-05-2024

Editor: Adrián Alejandro Vitón Castillo

ABSTRACT

Introduction: the exponential growth of data generation has led to an escalating concern for data privacy on a global scale. This work introduces a pioneering approach to address the often overlooked data privacy leakages associated with quasi-identifiers, leveraging artificial intelligence, machine learning and data correlation analysis as foundational tools. Traditional data privacy measures predominantly focus on anonymizing sensitive attributes and exact identifiers, leaving quasi-identifiers in their raw form, potentially exposing privacy vulnerabilities.
Objective: the primary objective of the presented work, is to anonymise the quasi-identifiers to enhance the overall data privacy preservation with minimal data utility degradation.
Methods: in this study, the authors propose the integration of ℓ-diversity data privacy algorithms with the OPTICS clustering technique and data correlation analysis to anonymize the quasi-identifiers.
Results: to assess its efficacy, the proposed approach is rigorously compared against benchmark algorithms. The datasets used are: Adult dataset and Heart Disease Dataset from the UCI machine learning repository. The comparative metrics are: Relative Distance, Information Loss, KL Divergence and Execution Time.
Conclusion: the comparative performance evaluation of the proposed methodology demonstrates its superiority over established benchmark techniques, positioning it as a promising solution for the requisite data privacy-preserving model. Moreover, this analysis underscores the imperative of integrating artificial intelligence (AI) methodologies into data privacy paradigms, emphasizing the necessity of such approaches in contemporary research and application domains.

Keywords: Data Privacy; Data Analysis; Data Processing; L-Diversity; Machine Learning; Clustering Algorithms.

RESUMEN

Introducción: el crecimiento exponencial de la generación de datos ha llevado a una creciente preocupación por la privacidad de los datos a escala global. Este trabajo presenta un enfoque pionero para abordar las fugas de privacidad de datos, a menudo pasadas por alto, asociadas a los cuasi-identificadores, aprovechando la inteligencia artificial, el aprendizaje automático y el análisis de correlación de datos como herramientas fundamentales. Las medidas tradicionales de protección de datos se centran principalmente en anonimizar los atributos sensibles y los identificadores exactos, dejando los cuasi-identificadores en su forma bruta, lo que expone potencialmente las vulnerabilidades de la privacidad.
Objetivo: el objetivo principal del trabajo presentado es anonimizar los cuasi-identificadores para mejorar la preservación general de la privacidad de los datos con una degradación mínima de la utilidad de los datos.

Métodos: en este estudio, los autores proponen la integración de algoritmos de privacidad de datos de ℓ-diversidad con la técnica de clustering OPTICS y el análisis de correlación de datos para anonimizar los cuasi-identificadores.
Resultados: para evaluar su eficacia, el enfoque propuesto se compara rigurosamente con algoritmos de referencia. Los conjuntos de datos utilizados son - Adult dataset y Heart Disease Dataset del repositorio de aprendizaje automático de la UCI. las métricas comparativas son: distancia relativa, pérdida de información, divergencia KL y tiempo de ejecución.
Conclusiones: la evaluación comparativa del rendimiento de la metodología propuesta demuestra su superioridad sobre las técnicas de referencia establecidas, posicionándola como una solución prometedora para el modelo de preservación de la privacidad de datos requerido. Además, este análisis subraya el imperativo de integrar metodologías de inteligencia artificial (IA) en los paradigmas de privacidad de datos, enfatizando la necesidad de tales enfoques en los dominios contemporáneos de investigación y aplicación.

Palabras clave: Privacidad de Datos; Análisis de Datos; Procesamiento de Datos; Diversidad L; Aprendizaje Automático; Algoritmos de Agrupación.

INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) increasingly dominate today’s tech landscape, necessitating vast databases for training and testing. Consequently, safeguarding personal data privacy is imperative. Data sanitation precedes its utilization to enhance model utility and machine training.

Data privacy, a vital aspect of data protection, ensures data integrity, confidentiality, and immutability. Compliance with regulations and robust mechanisms are essential for safeguarding sensitive information Priyank et al. (2016) Biswas et al. (2023). Data privacy differs from data security, focusing on authorized access, while the latter prevents hostile threats Gehrke et al. (2011) Kifer and Machanavajjhala (2011) Biswas et al. (2021).

Data can be categorized into Explicit identifiers, Quasi identifiers, and Sensitive attributes. Explicit identifiers reveal identity, while Quasi identifiers, when combined with additional data, pose privacy risks. Sensitive attributes include private information like salary and disease status. Explicit identifiers are hidden, sensitive attributes are protected using privacy algorithms, and Quasi identifiers, if overlooked, lead to unintended data privacy breaches.

There are many data-preserving algorithms researched and implemented, K-Anonymization: To achieve k-anonymity, the dataset must contain at least k records that share the same set of attributes that may be used to identify every individual Zheng et al. (2017)Sweeney (2002)Sweeney (2002)LeFevre et al. (2005). ℓ-diversity: By expanding the equivalence classes we developed with K-anonymity and masking the quasi-identifiers to include the confidential attributes in the record, ℓ-diversity seeks to expand on our privacy. According to the ℓ-diversity principle, the sensitive attribute has at least l “well-represented” values, then the data set is said to satisfy the ℓ-diversity criterion. ℓ-Diversity can be defined using different ways:

- Frequency ℓ-Diversity: if, for each equivalency class in the database, the relative frequency of each distinct sensitive value is not greater than 1/l, then the database can be considered anonymized using frequency ℓ-diversity.
- Entropy ℓ-Diversity: entropy of an equivalent class, E can be defined as:

$$-\sum_{E,s\in D(s)} p(E,s)\ log(p(E,s)) \qquad (1)$$

Where

$$p(E,s)=\frac{n(E,s)}{\sum_{w\in D(s)} n(w)}$$

That is, $p(E,s)$, the fraction of records in E that have the sensitive value s, is negated by the summation of s across the domain of the sensitive attribute of $p(E,s)log(p(E,s))$.

Despite its capabilities, the ℓ-diversity framework, a crucial component in data privacy, exhibits limitations as it does not consider the rarity associated with all sensitive values.

T- Closeness: an equivalency class is considered to have t-closeness if the difference between the distribution of a sensitive attribute inside it and the attribute distribution across the table is less than a threshold value, t.

Differential Privacy: differential Privacy ensures that if there exists two datasets say D_1 and D_2, one containing a particular information and the other without the information then, if a statistical query is executed over D_1 and D_2 then the probability of generation of a certain result is (almost) the same.

This paper presents a novel method for anonymizing quasi-identifiers to mitigate linkage, background-knowledge, and homogeneity attacks, thereby preventing inadvertent data privacy breaches. However, anonymizing all quasi-identifiers may significantly degrade data utility. To address these challenges, the proposed framework leverages data correlation analysis, clustering algorithms, and the ℓ-diversity data privacy preservation algorithm. Table 1 gives a systematic literature review of the related research works.

Organisation of the paper

The paper is structured into five sections to facilitate comprehensive understanding. Section 1 provides an Introduction, offering a brief overview of the research topic and its significance. In Section 2, Basic principles and theories pertinent to the study are elucidated to establish a foundational understanding. Section 3 outlines the Proposed Methodology, detailing the approach and techniques employed in the research. Section 4, Experimental Analysis, presents the findings and results derived from practical experimentation. Finally, Section 5 offers the Conclusion, summarizing key findings, implications, and avenues for future research, thus providing a holistic view of the study.

Table 1. Literature Review

S. No.	Research References	Key Points
1	Domingo-Ferrer and Mateo-Sanz (2002)	Used Micro-aggregation technique for control of data disclosure
2	Josep Domingo-Ferrer and Solanas (2007)	Micro-aggregation technique for used for disclosure control of multiple sensitive attributes.
3	Machanavajjhala et al. (2007)	Gave the official definition of ℓ-diversity algorithm Stated its need and importance
4	Hongwei Tian (2011)	Proposed Functional (τ,ℓ)-diversity that applied constraints over frequency of the sensitive attribute. Gave better results in terms of data utility & execution time when compared to the works of Machanavajjhala et al. (2007) & Gabriel Ghinita (2007)
5	Yuichi Sei and Ohsuga (2019)	Anonymization of Q.I.s using ℓ-diversity Adopted a probalistic approach Anonymization of all Q.I.s, lead to high degradation of data utility
6	Pooja Parameshwarappa (2020) and Pooja Parameshwarappa (2021)	Proposed multi-level clustering approach Different methods of clustering were used
7	Ren (2021)	Proposed privacy preservation of IoT generated data by anonymization of Q.I.s
8	Brijesh B. Mehta (2022)	Proposed Improved Scalable ℓ-diversity algorithm (ImSLD) Used Map-Reduce technology to handle big data privacy preservation
9	Dunbo Cai and Huang (2022)	Proposed privacy preservation of Q.I.s in a free text model Absence of benchmark datasets for comparison of results Scalability issues can hinder its practical implementation

Basic Principles and Theories

Mutual Information Correlation

It deals with quantification of correlation between random variables. It is different from linear correlation as it deals with non-linear correlation. Mutual Information is a transformation of correlation. Thus, the Mutual correlation between random variables X and Y is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p_{(X,Y)}(x,y)\, log \frac{p_{(X,Y)}(x,y)}{p_{(X)}(x) p_{(Y)}(y)} \qquad (2)$$

Here $p_{(X)}$ and $p_{(Y)}$ are the marginal probability density functions, and $p_{(X,Y)}$ is the joint probability density function. Whereas to compute the mutual information for continuous random variables, the integrals must replace the summations.

$$I(X;Y) = \int_Y \int_X p_{(X,Y)}(x,y)\, log \frac{p_{(X,Y)}(x,y)}{p_{(X)}(x) p_{(Y)}(y)}\, dx\, dy \qquad (3)$$

https://doi.org/10.56294/dm2024287

Relative Distance

It is a ratio of the value difference by the maximum value among the two. Let X be the original value of an attribute, and Y be the value of the same attribute after applying an algorithm. Then, the relative distance gets calculated as:

$$RD(X,Y) = \frac{|X-Y|}{max(X,Y)} \qquad (4)$$

The more the value of R.D., the more is the data privacy and lesser is the data utility.

Information Loss

It is a metric used to estimate the amount of information lost in the process of applying an algorithm. Here, this algorithm refers to a data privacy algorithm. Let X be an attribute of dataset D before applying any data privacy algorithm. Let Mean be the mathematical mean of X. Now, after applying the data privacy algorithm, attribute X becomes Y. Let Mean' be the mathematical mean of the attribute Y. Then, information loss between X and Y gets calculated as:

$$Information\ Loss(X,Y) = \frac{|Mean - Mean'|}{(Mean + Mean')} \qquad (5)$$

A higher value of I.L. corresponds to a higher data privacy value. However, this enhancement in privacy is accompanied by a decrease in the data utility measure.

KL Divergence

KL divergence quantifies how much one probability distribution differs from another probability distribution. Let dataset P transform to dataset Q upon application of a data privacy algorithm. Then, the KL divergence gets calculated using the following formula:

$$KL(P||Q) = -\sum_{x=1}^{all\ attributes} P(x) * \frac{log\ P(x)}{log\ Q(x)} \qquad (6)$$

Where P(x) and Q(x) are the probability distributions of attributes in datasets P and Q, respectively.

A low KL value is desirable as the application of a data privacy preservation algorithm must not change the original probability distribution of the dataset.

PROPOSED METHODOLOGY

Description of the proposed workflow

The methodology proposed in this work utilises the following concepts - Data correlation analysis using the Mutual Information Correlation analysis technique, machine learning classification algorithms for classifying attributes into various categories, and application of ℓ-diversity algorithm to ensure data privacy. The concepts are arranged into the following steps:

Selection of attributes for anonymisation: let D be the original dataset containing m attributes (A_1,A_2,A_3,......,A_m) and n no. of rows (r_1,r_2,r_3,.....,r_n). The attributes of D can be classified as: Sensitive Attributes (S), Exact Identifiers(EI) and Quasi Identifiers (QI). In the first step, we further try to segregate the Quasi identifiers into two categories: Sensitive Quasi Identifiers (SQI) and Non-Sensitive Quasi Identifiers (NSQI). The classification is done with the help of the mutual information correlation analysis technique (MIC). The Data Correlation Co-efficient values between every QI and sensitive attribute is calculated using MIC technique. Then, for all the QIs, we check the following condition:

$$DataCorrelationCoefficient(Q.I.,S) \geq \eta \qquad (7)$$

If the condition evaluates as True, then the Q.I. is classified as S.Q.I else it gets classified as N.S.Q.I figure 1 represents this step diagrammatically in detail.

Generation of the OPTICS' algorithm: in this step, we use the MIC algorithm to calculate the Data Correlation Coefficient Matrix corresponding to the rows. Then, use the generated data correlation coefficients with the conventional OPTICS clustering algorithm to generate OPTICS' algorithm. We have used the OPTICS algorithm in this step because it efficiently handles outliers while clustering.

OPTICS’ algorithm

Instead of using Euclidean distance to calculate reachability distance, we have proposed using Correlated Euclidean Distance (denoted as CED). Then, we calculate CED between objects x_1 and x_2 using the following mathematical formula:

$$CED(x_1, x_2) = EuclideanDistance(x_1, x_2) * \frac{1}{DataCorrelationCoefficient(x_1, x_2)} \tag{8}$$

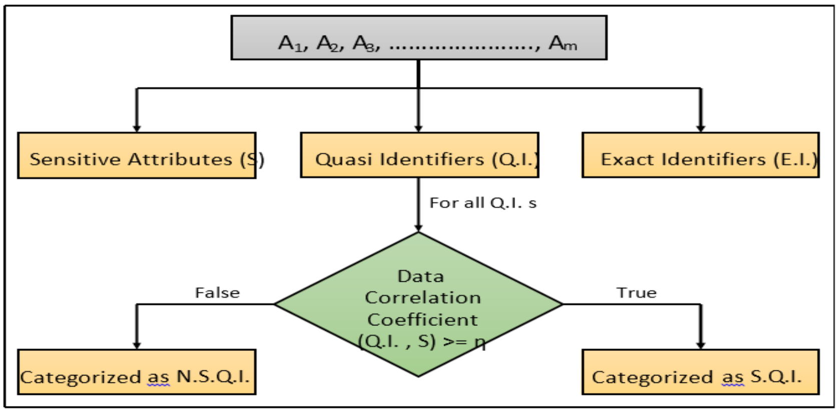

Figure 1. Diagrammatic description of Step 1

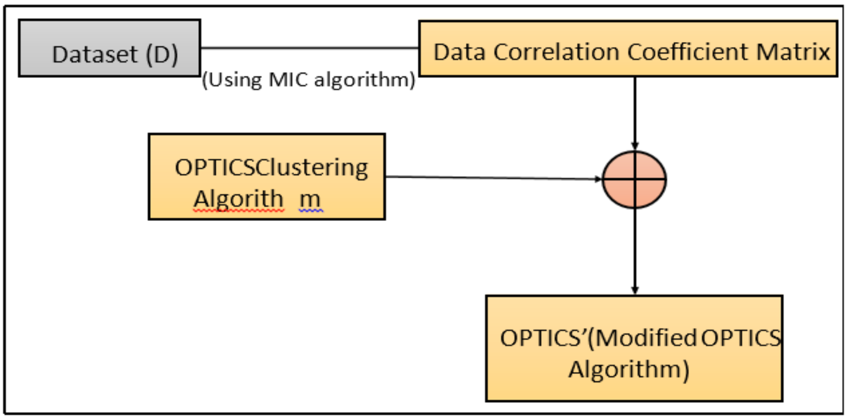
Dataset (D)
(Using MIC algorithm)
Data Correlation Coefficient Matrix
OPTICSClustering Algorith m
OPTICS’(Modified OPTICS Algorithm)

Figure 2. Diagrammatic description of Step 2

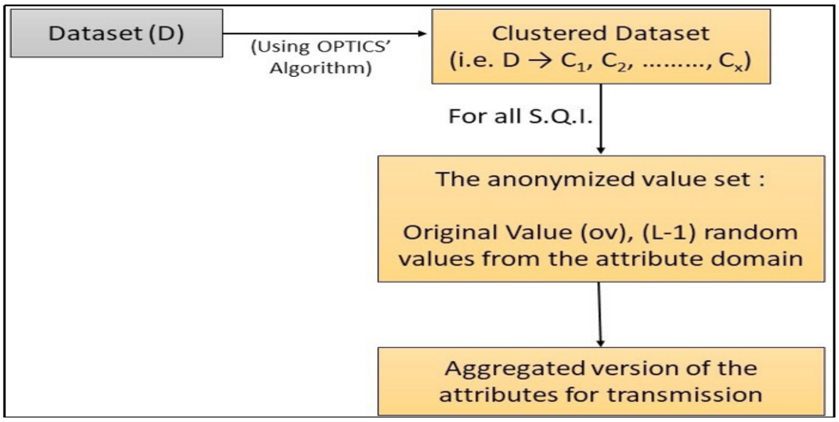
Dataset (D)
(Using OPTICS’ Algorithm)
Clustered Dataset (i.e. D → C1, C2, ..., Cx)
For all S.Q.I.
The anonymized value set :
Original Value (ov), (L-1) random values from the attribute domain
Aggregated version of the attributes for transmission

Figure 3. Diagrammatic description of Step 3 and 4

Clustering the dataset: using the OPTICS’ algorithm, now perform the clustering of dataset D to generate a partitioned dataset.

In the initial step of the process, a set of Sensitive Quasi Identifiers (S.Q.I.) is generated. Subsequently, to uphold data privacy, ℓ-diversity is employed as a pivotal mechanism for anonymization. This involves the generation of anonymized value sets for each S.Q.I., ensuring that the resulting data maintains a diverse representation of sensitive attributes. By performing ℓ-diversity on the generated S.Q.I., the study seeks to strike a balance between data privacy and analytical utility, contributing to the broader discourse on effective privacy-preserving methodologies in data processing.

Let ov be the original value corresponding to cell (r_i,$S.Q.I_j$) of D, then the Anonymised value set corresponding to ov will be ov , (L-1) random values from attribute domain where attribute domain refers to the domain of the $S.Q.I._j$ within the cluster to which r_i belongs.

Prepare the aggregated version for transmission. Figure 3 is a diagrammatic representation of Step 3, Step 4 and Step 5.

Table 2. Statistical Features of ADULT Dataset

Attribute Name	Mean	Median	Standard Deviation	Minimum Value	Maximum Value	Mode
Age	38,581	37,0	13,640	17	90	-
Education-num	10,080	10,0	2,572	1	16	-
Hours-per-week	40,437	40,0	12,347	1	99	-
Capital-gain	1077,64	0,0	7385,29	0	99999	-
Capital-loss	87,303	0,0	402,96	0	4356	-
Workclass	-	-	-	-	-	Private
Education	-	-	-	-	-	HS-grad
Occupation	-	-	-	-	-	Prof-specialty
Marital-status	-	-	-	-	-	Married-civ-spouse
Relationship	-	-	-	-	-	Husband
Race	-	-	-	-	-	White
Sex	-	-	-	-	-	Male
Native-country	-	-	-	-	-	United-States

Table 3. Statistical Features of HEART-DISEASE Dataset

Attribute Name	Mean	Median	Standard Deviation	Minimum Value	Maximum Value
Age	54,542	56,0	9,049	29	77
Sex	0,6767	1,0	0,468	0	1
Cp	3,158	3,0	0,964	1	4
Trestbps	131,693	130,0	17,762	94	200
Chol	247,350	243,0	51,997	126	564
Fbs	0,144	0,0	0,352	0	1
Restecg	0,996	1,0	0,994	0	2
Thalach	149,599	153,0	22,941	71	202
Exang	0,326	0,0	0,469	0	1
Oldpeak	1,055	0,8	1,166	0,0	6,2
Slope	1,602	2,0	0,618	1	3
Ca	0,676	0,0	0,938	0	3
Thal	4,730	3,0	1,938	3	7
Goal	0,946	0,0	1,234	0	4

Experimental Analysis

Dataset Description

In the study that is being given, we conducted tests to analyze the suggested strategy using the Adult and Heart-Disease Dataset from the UCI Machine Learning Repository.

The Adult dataset gives information about an individual's annual income resulting from various factors. Intuitively, the individual's education level, age, gender, occupation, and other factors influence it. The aforementioned is a widely cited KNN dataset. It has 48862 rows, and we have used all the columns with 'Income' as the sensitive attribute.

The Heart-Disease dataset uses 13 major parameters to detect the presence of Heart disease in a patient. It has 303 rows and we have used the 13 prominent features. Table 2 and table 3 display the statistical features of the attributes of the ADULT and HEART-DISEASE datasets, respectively.

Algorithms Compared in the Experiments

For the evaluation of our proposed methods using experimental simulations, we have used the following algorithms:

LDPA: the method was proposed in Yuichi Sei and Ohsuga (2019) and used a probabilistic approach to implement ℓ-diversity. This is referred to as LDPA in our work and stands for ℓ-Diversity using the Probabilistic Approach.

MC-ℓ-MDAV: this is one of the proposed algorithms in the work Pooja Parameshwarappa (2021). It uses multi-level clustering with Euclidean Distance and backtracking to cluster the activity sequences Pooja Parameshwarappa (2021).

MC-ℓ- VMDAV: it is another algorithm proposed in Pooja Parameshwarappa (2021). It clusters the activity sequences by using multi-level clustering with Euclidean Distance and VMDAV Pooja Parameshwarappa (2021).

Standard ℓ- diversity algorithm: this is the standard ℓ-diversity algorithm.

LDCML: this algorithm proposed in the current work stands for ℓ-Diversity using correlation analysis and machine learning.

EXPERIMENTAL RESULTS

This section provides a comprehensive exposition of the results obtained through an experimental analysis of the algorithms discussed in the preceding subsection. The focus of this analysis encompasses a detailed examination of the performance, efficacy, and characteristics of the algorithms under scrutiny. By rigorously evaluating the outcomes derived from the implemented algorithms, we aim to elucidate their strengths, limitations, and overall suitability for the specific tasks outlined in the preceding sections. The comparative analysis is done based on the following factors - Relative Distance, Information Loss, KL Divergence value and Execution Time.

The first set of observations show that the LDPA algorithm has the highest value of Relative Distance and thus will have the lowest data utility. The Relative Distance values of the MC-ℓ-MDAV, MC-ℓ-VMDAV and LDCML algorithms are higher than the standard ℓ-diversity algorithm. This observation is due to the standard algorithm's low data privacy guarantee. The LDCML algorithm proposes anonymisation of quasi-identifiers along with the sensitive attributes with nearly equal values of relative distance when compared to the MC-ℓ-MDAV and MC-ℓ-VMDAV algorithms.

Nearly the same trend can be observed for the Information Loss Values for various mentioned algorithms. The trend's reasons are similar to those mentioned in the above paragraph. Thus, by analysing the five algorithms in terms of relative distance and information loss, we have derived the following conclusion regarding the Data Utility:

DataUtility(Standardℓ) > DataUtility(MC−ℓ−MDAV) ≈ DataUtility(MC− ℓ − V MDAV) ≈ DataUtility(LDCML) > DataUtility(LDPA).

Table 4. Values of Various Comparative Metric

S.No.	Method	Relative Distance	Information Loss	KL Divergence	Execution Time
1	LDPA	0,45	0,39	0,49	231,70
2	MC-ℓ-MDAV	0,33	0,30	0,42	202,79
3	MC-ℓ-VMDAV	0,18	0,28	0,36	42,53
4	Standard ℓ-diversity	0,22	0,16	0,20	16,17
5	LDCML	0,24	0,23	0,24	26,38

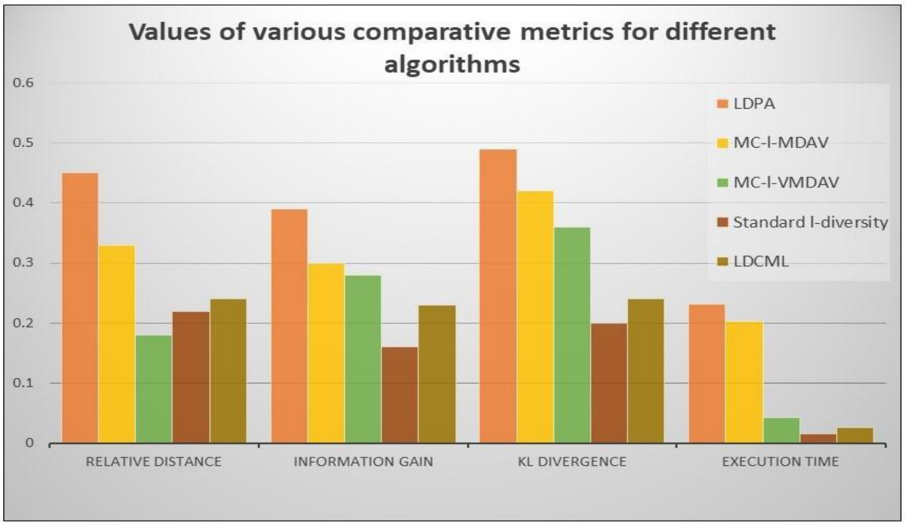

Figure 4. Values of various comparative metrics for different algorithms

Using KL divergence, we can measure the probability distribution of data before and after the application of the proposed algorithm. The difference must be small, as the distribution must not change with the application of the algorithm. If the distribution changes by a larger magnitude, it implies that any intruder or unauthorised entity can gain information by using this difference. The observations recorded with respect to KL DIvergence show that the standard ℓ-diversity algorithm has the least value of KL Divergence followed by our proposed method, i.e., the LDCML method. The values of KL divergence for the rest of the methods are higher than the LDCML method.

In terms of Execution Time, the Standard ℓ-diversity algorithm takes the least time for execution due to its ease of application. Among the rest of the algorithms, the proposed LDCML algorithm has the lowest execution time.

Table 4 and figure 4 depict the results in a tabular format and graphically, respectively for easy interpretation.

CONCLUSION

The presented work introduces a groundbreaking approach, LDCML, for anonymizing sensitive quasi-identifiers and attributes utilizing artificial intelligence tools to ensure comprehensive data privacy in any given dataset. The proposed algorithm aims to strike a balance between ensuring data privacy and maintaining the necessary level of data utility. The evaluation of the algorithm's effectiveness involves a comprehensive analysis, considering Information Loss, Relative Distance, KL Divergence, and Execution Time.

To assess the proposed LDCML algorithm's performance, comparative analyses are conducted against established algorithms, namely LDPA, MC-ℓ-MDAV, and MCℓ-VMDAV. Results indicate that LDCML exhibits comparable data utility values to these benchmark algorithms. Moreover, LDCML outperforms its counterparts in terms of KL Divergence, establishing its supremacy in minimizing information loss.

Furthermore, the execution time analysis reveals that LDCML demonstrates a more efficient processing time when compared to LDPA, MC-ℓ-MDAV, and MC-ℓVMDAV algorithms. This signifies the proposed algorithm's efficiency in anonymizing quasi-identifiers and sensitive attributes while maintaining a superior level of execution speed.

These findings position LDCML as a promising solution for privacy-conscious data anonymization processes and profess that the integration of AI technologies emerges as a pivotal strategy in the evolving landscape of data protection, emphasizing the growing significance of artificial intelligence in safeguarding sensitive information.

Limitations and Future Scope

The study at hand does not include a comprehensive analysis of the computational complexity associated with the proposed methodology. Additionally, future investigations could explore the integration of advanced clustering algorithms and the incorporation of fuzzy logic systems to enhance the depth and scope of research in this domain.

Declarations

Availability of Data and Materials

The datasets used are publicly available on the UCI Machine Learning Repository. The source code and other associated data will be made available by the authors on requests.

REFERENCES

1. Priyank J, Manasi G, Nilay K. Big data privacy: a technological perspective and review. Journal of Big Data. 2016 03; https://doi.org/10.1186/s40537-016-0059-y.

2. Biswas S, Fole A, Khare N, Agrawal P. Enhancing correlated big data privacy using differential privacy and machine learning. Journal of Big Data. 2023 March;10.

3. Gehrke J, Lui E, Pass R. Towards Privacy for Social Networks: A Zero-Knowledge Based Definition of Privacy. In: Theory of Cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 432–449.

4. Kifer D, Machanavajjhala A. No Free Lunch in Data Privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. SIGMOD '11. New York, NY, USA: Association for Computing Machinery; 2011. p. 193–204. Available from: https://doi.org/10.1145/1989323.1989345.

5. Biswas S, Khare N, Agrawal P, Jain P. Machine learning concepts for correlated big data privacy. Journal of Big Data. 2021 december;8.

6. Zheng L, Yue H, Zhaoxuan L, Pan X, Wu M, Yang F. k-Anonymity Location Privacy Algorithm Based on Clustering. IEEE Access. 2017 12;PP:1–1. https://doi.org/10.1109/ACCESS.2017.2780111.

7. Sweeney L. K-Anonymity: A Model for Protecting Privacy. Int J Uncertain Fuzziness Knowl-Based Syst. 2002 oct;10(5):557–570. https://doi.org/10.1142/S0218488502001648.

8. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient Full-Domain KAnonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. SIGMOD '05. New York, NY, USA: Association for Computing Machinery; 2005. p. 49–60. Available from: https://doi.org/10.1145/1066157.1066164.

9. Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transaction of Knowledge Data Engineering, 14. 2002;p. 189–201.

10. Josep Domingo-Ferrer FS, Solanas A. Microaggregation heuristics for p-sensitive k anonymity; 2007.

11. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-Diversity: Privacy beyond k-Anonymity. ACM Trans Knowl Discov Data. 2007 Mar;1(1):3–es. https://doi.org/10.1145/1217299.1217302.

12. Hongwei Tian WZ. Extending l-diversity to generalize sensitive data. Data & Knowledge Engineering, 70. 2011;p. 101–126.

13. Gabriel Ghinita PKNM Panagiotis Karras. Fast data anonymization with low information loss. International Conference on Very Large Data Bases. 2007;p. 758–769.

14. Yuichi Sei TT Hiroshi Okumura, Ohsuga A. Anonymization of Sensitive QuasiIdentifiers for l- Diversity and t-closeness. vol. 16; 2019.

15. Pooja Parameshwarappa GK Zhiyuan Chen. An effective and computationally efficient approach for anonymizing large-scale physical activity data: multi-level clusteringbased anonymization. International Journal of Information Security and Privacy, IJISP. 2020;

16. Pooja Parameshwarappa GK Zhiyuan Chen. Anonymization of Daily Activity Data by using l-diversity Privacy Model. ACM Trans Manage Inf Syst 12. 2021 May;.

17. Ren TXDJea W. Privacy Enhancing Techniques in the Internet of Things Using Data Anonymisation. Inf Syst Front. 2021 06;30. https://doi.org/10.1007/s10796-021-10116-w.

18. Brijesh B Mehta UPR. Improved l-diversity: Scalable anonymization approach for privacy preserving big data publishing. Journal of King Saud University - Computer and Information Sciences. 2022;.

19. Dunbo Cai DX Ling Qian, Huang Z. Towards a Free Text Dataset for Hiding Quasi-Identifiers; 2022. p. 1-6.

20. Sweeney L. Achieving K-Anonymity Privacy Protection Using Generalization and Suppression. Int J Uncertain Fuzziness Knowl-Based Syst. 2002 oct;10(5):571–588. https://doi.org/10.1142/S021848850200165X.

21. Han Jian-min CTT, Hui-Qun Y. An improved V-MDAV algorithm for l-diversity; 2008. p. 733–739.

22. Jain P, Gyanchandani M, Khare N. Differential privacy: its technological prescriptive using big data. Journal of Big Data. 2018 04;5. https://doi.org/10.1186/s40537-018-0124-9.

23. Priyank J, Manasi G, Nilay K. Enhanced Secured Map Reduce layer for Big Data privacy and security. Journal of Big Data. 2019 06; https://doi.org/10.1186/s40537-019-0193-4.

24. Chen J, Ma H, Zhao D, Liu L. Correlated Differential Privacy Protection for Mobile Crowdsensing. IEEE Transactions on Big Data. 2021 oct;7(04). https://doi.org/10.1109/TBDATA.2017.2777862.

25. Yang Xinyu RXYW Wang Teng. Survey on Improving Data Utility in Differentially Private Sequential Data Publishing. IEEE Transactions on Big Data. 2017; https://doi.org/10.1109/TBDATA.2017.2715334.

26. Fei F, Li S, Dai H, Hu C, Dou W, Ni Q. A K-Anonymity Based Schema for Location Privacy Preservation. IEEE Transactions on Sustainable Computing. 2019;4(2):156– 167. https://doi.org/10.1109/TSUSC.2017.2733018.

27. El Ouazzani Z, El Bakkali H. A New Technique Ensuring Privacy in Big Data. Procedia Comput Sci. 2018 may;127(C):52–59. https://doi.org/10.1016/j.procs.2018.01.097.

28. Djoudi M, LyndaKacha, AbdelhafidZitouni. KAB : A new k-anonymity approach based on black hole algorithm. 2021;.

29. Arava K, Lingamgunta S. Adaptive K-anonymity approach for privacy preserving in cloud. 2019;p. 1-8.

30. Raghuraj G, Naikodi DC, L S. K-Anonymization-Based Temporal Attack Risk Detection Using Machine Learning Paradigms. Journal of Circuits, Systems and Computers. 2020 06;30. https://doi.org/10.1142/S021812662150050X.

CONFLICT OF INTEREST

No conflict of interest is associated.

FUNDING DETAILS

Funding details are not applicable.

AUTHORSHIP CONTRIBUTIONS

Conceptualization: Sreemoyee Biswas.
Data curation: Sreemoyee Biswas and Priyank Jain.
Formal analysis: Sreemoyee Biswas, Vrashti Nagar, Priyank Jain, Nilay Khare.
Methodology: Sreemoyee Biswas.
Supervision: Nilay Khare, Pragati Agrawal.
Validation: Priyank Jain.
Visualization: Sreemoyee Biswas, Vrashti Nagar, Priyank Jain.
Writing - original draft: Sreemoyee Biswas, Vrashti Nagar.
Writing - review and editing: Priyank Jain, Nilay Khare, Pragati Agrawal.