2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Generating Privacy-Preserving Data without Compromising Analytical Utility

Xiaopeng Luo[1], Shanlin Feng[1], Yunlin Liu[1], and Taoting Xiao[2] (✉)

Guangzhou Huali College, Guangzhou, China
gordonlok@foxmail.com, 3502396471@qq.com, 2126227224@qq.com
Guangzhou College of Commerce, Guangzhou, China
geekarina96@gmail.com

Abstract. The exponential progress in machine learning has created an unprecedented need for diverse and high-quality datasets. However, efficiently extracting meaningful insights from confidential information remains a significant hurdle, as improper handling of sensitive user data can compromise privacy—a cornerstone of reliable data-driven systems. Modern data collection practices, fueled by ubiquitous digital devices, frequently generate mixed-format datasets combining continuous numerical variables and discrete categorical attributes. While such heterogeneous data enhances analytical depth by revealing complex feature-label relationships, its usefulness is often undermined by non-informative variables that contribute more to noise than to predictive accuracy. To overcome these limitations, we present a privacy-aware synthetic data generation framework that dynamically evaluates feature relevance, even in low-data regimes. Leveraging rigorous differential privacy mechanisms, our approach injects mathematically structured noise to obfuscate personal identifiers while preserving dataset integrity. The synthesized output demonstrates enhanced statistical diversity without sacrificing analytical utility. Comprehensive benchmarking confirms that classifiers trained on our synthesized data achieve superior performance compared to conventional methods, all while maintaining strict privacy guarantees. This dual advantage addresses critical gaps in secure machine learning pipelines.

Keywords: Mixed data, Feature selection, Feature ranking, Privacy-preserving.

1 Introduction

With the rapid development of machine learning in recent years, the scale of models has been gradually increasing, and models have been widely adopted across various industries. However, the current models have increasingly high requirements for real-world datasets[25]. Collecting data and analyzing data has incurred serious privacy issues since such data contain various sensitive information of users. Even worse is that driven by advanced data fusion and analysis techniques[48], the real-world data of users are more vulnerable to attack and disclosure in the big data era[22,23].

Synthetic data can provide anon-real dataset similar to the original data and avoid exposing users' sensitive information while collecting and analyzing data[27,28,12,29].

Additionally, classification algorithms are practical and effective tools in data analysis. It is a challenging problem to capture valuable information in the trained classifier. Their accuracy heavily relies on the features that describe the training data. In practice, the increase in types of data collection devices results in heterogeneous, correlated, redundant, and irrelevant label features. Ir- relevant features may not significantly impact the analysis results, but they can be exploited through association attacks to infer user information, which poses a security threat to users [43,16,38].

However, irrelevant features can be not as useless as expected. We can appropriately utilize these features to expand the size of the dataset while safeguarding the user information present in the original data without causing significant loss of information. In deep learning, adding Gaussian noise to the input data during training is common to enhance models' robustness and generalization capabilities[20,21,24]. This technique is known as data augmentation. Generative Adversarial Networks (GANs) can also generate data and introduce noise to the input data, improving the model's robustness [46,32]. However, such deep learning methods require substantial amounts of data, ample computational re- sources, and a strict definition of privacy.

Based on the above discussion, it is essential to identify the features in the mixed data unrelated to the labels to synthesize privacy-protected data from a limited amount of original data. By altering the values of these features, new data can be generated, effectively expanding the dataset. In differential privacy [31], adding unbiased noise ensures that data privacy is mathematically constrained while preserving good data utility without affecting the data structure. Therefore, this paper analyzes the structure and relationships of heterogeneous features to perform feature selection from a homogeneous perspective.

Hence, this study leverages the characteristics of graph structures in data representation to compute the correlation between heterogeneous features and labels. Furthermore, suitable differential privacy perturbation methods are employed for different types of features. Extensive experiments on the benchmark datasets have demonstrated the superiority of the proposed approach. The main contributions of this paper are summarized in the following.

- Our proposed feature selection method utilizes the graph structure of features to calculate the correlation between the labels and heterogeneous features. We eliminate the requirement of pre-allocated parameters for correlation estimation.
- We propose a heterogeneous differential privacy perturbation framework that applies to different features. This framework balances privacy preservation and data utility. It ensures that private information is protected while maintaining the usefulness of the data.
- Our method has been applied to public and personal datasets, including music emotion classification, breast cancer, and credit card data. In particular, we generated synthetic datasets using our method and utilized them to train two commonly used classifiers. The proposed method is competent to synthetic data and shows its efficacy

in comparison with the original data.

2 Related Work

This section overviews the representation learning methods, data synthesis and differential privacy.

2.1 Representation Learning Methods

Representation learning methods include the conventional approaches that learn the importance of a whole attribute. Advanced methods [37,39] first informatively represent categorical values, then learn the represented inter-value distances and attribute importance. However, they are still based on hand-crafted encoding and measures. Most recently, nested cluster distribution learning approaches[8,10,26,41], ordinal attribute value relationship mining methods[26,35,44], and distributed clustering algorithms[6,15,40,48] have been proposed. Since they focus on addressing the other complexity in the clustering field, they are relatively incompetent in solving our focused metric space and attribute subspace learning problems.

2.2 Data Synthesis

In recent years, Generative Adversarial Networks (GANs) have been widely applied in data synthesis, specifically for generating tabular data. Many studies have improved the architecture, loss functions, and optimization strategies of GANs to address specific limitations. For example, improved GAN architectures such as WGAN [3] and WGAN-GP [13] have significantly progressed in training stability and convergence time. PacGAN [16] addresses mode collapse issues in traditional GAN architectures. However, challenges still exist in handling data heterogeneity, which includes numerical and categorical features.

CTGAN [30] is an advanced method that has progress. It can synthesize heterogeneous tabular data. However, CTGAN requires extensive experimentation for optimizing hyperparameters, especially for datasets with different features. For small datasets, insufficient records would make CTGAN perform badly for training and convergence. The training and convergence processes may also demand significant computational resources and time, especially for large datasets.

Regarding synthesizing private data, probabilistic graphical models have become the most promising methods for privately generating synthetic datasets. PrivBayes [34] utilizes Bayesian networks to determine the network structure and obtain noise margins for the conditional probability distributions of each node. Another PGM-based method, based on Markov Random Fields [34], also performs well. These methods offer advantages in terms of computational efficiency or empirical performance compared to non-PGM-based synthesis methods [1,5]. However, once a graph structure is fixed, they impose assumptions of conditional independence that may not hold in the dataset. The PrivSyn [42] method makes weaker assumptions about the conditional independence between features and attempts to capture the relevant relationships in the dataset. However, these methods, including PGM and PrivBayes, need to adequately discuss leveraging the heterogeneous information in tabular data and quantifying the relationships between various features and labels.

Furthermore, when dealing with datasets containing many features, these methods require significant storage space to store the graph structure and search for the optimal one. Computing the correlations between features also consumes a substantial amount of computational resources. In this study, we focus on two frequently utilized techniques in differential privacy: Information Gain and Chi-square. [7,2].

2.3 Differential Privacy

Differential privacy refers to the concept where a trusted data curator collects data from individual users, processes it in a way that satisfies the principles of differential privacy, and then releases the results. In essence, the concept of differential privacy requires that the influence of any individual element in the dataseton the output is as small as possible.

Definition 1. Differential Privacy: An algorithm ***A*** satisfies (ϵ, δ)-differential privacy ((ϵ,δ)-DP), where $\epsilon > 0$, $\delta \geq 0$, if and only if for any two neighboring datasets D and D' , we have:

$$\forall S \subseteq \mathrm{Range}(A) \text{: } \Pr[A(D) \in S] \leq e^{\epsilon} \Pr[\mathrm{A}(\mathrm{D}_0) \in S] + \delta,$$

where Range(A) denotes the set of all possible outputs of the algorithm ***A***.

Differential privacy can be applied to various data processing scenarios, including data mining, machine learning, and statistical analysis. It can protect personal identities, sensitive features, and private information while allowing meaningful data analysis and inference.

The concept and techniques of differential privacy have been extensively researched and applied in academia and industry. Many mechanisms and algorithms for differential privacy have been proposed, such as the Laplace mechanism, exponential mechanism, and confusion matrices.

Furthermore, differential privacy has been incorporated into some privacy laws and regulations, becoming an important standard for privacy protection.

In summary, differential privacy provides a powerful method for privacy protection in individual data processing. It not only safeguards individuals' privacy but also supports meaningful data analysis. It seeks to balance privacy protection and data utility, supporting data-driven societal and technological advancements.

3 Proposed Method

In this section, we first define the symbols representing data components, heterogeneity measures, feature dependencies, and the differential privacy algorithm used for data synthesis. Then, we construct the spatial structure for heterogeneous features and propose a metric that measures the correlation between features and labels. Finally, we present a comprehensive data synthesis frame- work.

3.1 Problem Formulation

In most real-world datasets, variations in feature values have different impacts on data labels. Each feature type has a unique graph structure representing its feature values.Our objective is to analyze the impact of attribute changes on classification while preserving the graph structure of the original data attributes. We aim to minimize the influence of subsequent noise addition on the original labels. Additionally, we strive to maintain unbiased mean estimation during the noise addition process, ensuring that the overall distribution of the dataset remains unbiased.

A heterogeneous feature dataset $\boldsymbol{D}$ can be represented as a tuple $\boldsymbol{D} = \langle X, F, V \rangle$, where X= $\{x_i \mid i = 1, 2, ..., n\}$ represents n users or data samples. $F = \{f_r \mid r = 1, 2, ..., d\}$ is the feature set composed of d features, including d_n nominal, do ordinal, and d_u numerical features. $V = \{V_r \mid r = 1, 2, ..., d\}$ is the collection of unique value sets of each feature, where $V_r = \{v_{r1}, v_{r2}, ..., v_{rk}\}$ is the unique value set of f_r . Compared to nominal data, numerical data and ordinal data have exhibited a progressive relationship, where $v_{r1} < v_{r2} < ... < v_{rk}$.

Therefore, we construct separate graph structures for the three data types, as shown in Figure 1.

We can observe that ordinal data is similar to numerical data. If we assume there is ordinal data with infinite feature values, it can be considered a special numerical data type. Similarly, ordinal data and nominal data also share similarities. In ordinal data, two feature values with a large span may exhibit contradictory characteristics, similar to nominal data. For example, in an ordinal feature describing breast cancer, the menstrual period and cancer occurrence have four feature values: ⟨No menstruation, menstruation under 40, menstruation above 40, menopause ⟩. "No menstruation" and "menopause" can be viewed as special nominal data.

We establish bins with the same number of labels C or numerical data, using equal-frequency partitioning. The bins' boundary values, denoted as M, are determined as $M = C + 1$. Subsequently, we construct node matrix Ur for three different types of features as follows: $U_r = \{\mathrm{u_{ri}} \mid i = 1, 2, \ldots., m\}$. It implies

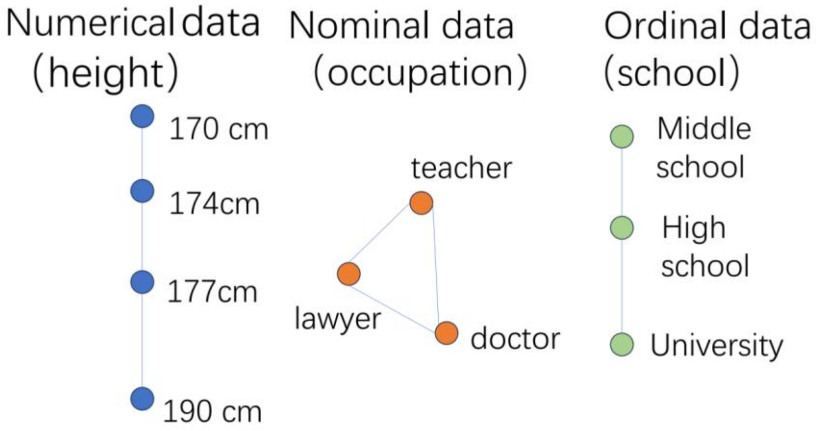

Fig. 1. Numerical and ordinal data have a progressive relationship between values, but numerical data can use Euclidean distance to express dissimilarity

that the boundary value, obtained from binning the numerical data, is used as anode. Similarly, the feature values are directly used as nodes for nominal and ordinal data. Specifically, in the case of numerical data, each data from the original dataset is assigned to the nearest node.For each label in the label set $C = \{c_j \mid j = 1, 2, ..., t\}$, we define the U_j^r of each label value C_j in feature F as follows:

$$U_j^r = [p(c_j|f_1), p(c_j|f_2), \ldots, p(c_j|f_d)],$$

The matrix U_r stores the probabilities of label values c_j appearing, given each feature value in feature r.When there are two label values, g and h, we define ΔU_{gh}^r, which indicates how the nodes corresponding to labels g and h on feature r undergo graph transformations and the differences that require offset transmission:

$$\Delta \mathrm{U}_{\mathrm{gh}}^{\mathrm{r}} = \mathrm{U}_{\mathrm{g}}^{\mathrm{r}} - \mathrm{U}_{\mathrm{h}}^{\mathrm{r}} \ ,$$

Simultaneously, we construct the vector $Cost_{\Delta \mathrm{U}_{\mathrm{gh}}^{\mathrm{r}}}^r$ to store the minimum cost for transmitting node transformations [36]:

$$Cost_{\Delta \mathrm{U}_{gh}^{\mathrm{r}}}^r = [Dis(v_{rp}, v_{rq})_1, \ldots, Dis(v_{rp}, v_{rq})_d].$$

The cost is calculated based on the extent of dissimilarity between the feature values of each feature. The calculation method for dissimilarity in different types of features is defined as follows:

$$Dis(v_{rp} - v_{rq}) = \begin{cases} |v_{rp} - v_{rq}| & , f_r \in \ numerical \\ 1 & , f_r \in \ nominal \\ |p - q| & , f_r \in \ ordinal \end{cases} \tag{1}$$

Obtaining $\mathrm{Cost}_{\Delta \mathrm{U}_{\mathrm{gh}}^{\mathrm{r}}}^{\mathrm{r}}$ will become a straightforward optimization problem in operations research. We can define the following method to calculate whether we can distinguish between label g and label h well in feature f_r:

$$\emptyset^{\mathrm{r}}(\mathrm{g},\mathrm{h}) = \left|\mathrm{U}_{\mathrm{g}}^{\mathrm{r}} - \mathrm{U}_{\mathrm{h}}^{\mathrm{r}}\right| \cdot \mathrm{Cost}_{\Delta \mathrm{U}_{\mathrm{gh}}^{\mathrm{r}}}^{\mathrm{r}} \tag{2}$$

Finally, we define the calculation method for the correlation between heterogeneous features and labels as follows:

$$w(f_r, \mathrm{Label}) = \frac{\sum_{g=1}^{t-1}\sum_{h=g+1}^{t} \emptyset^{r}(g,h)}{t-1} \tag{3}$$

The complete process example of *w* (*f*$_r$, Label) calculation is shown in Figure 2. The current feature can effectively distinguish between labels if *w*(*f*$_r$, Label) is large. Conversely, if *w*(*f*$_r$, Label) is small, it suggests that utilizing this feature may not contribute significantly to distinguishing labels in the classification task.

3.2 Perturbation mechanism

After obtaining the weights *w*(*f*$_r$, Label), we apply the perturbation mechanism to the feature with the minimum weight. The purpose is to prevent users from being attacked by adversaries exploiting background knowledge or differential attacks to obtain sensitive information. Additionally, since the feature with the minimum weight has a relatively smaller impact on the classification task, perturbing this feature ensures that the synthetic data can retain the utility of the original data. We will implement corresponding perturbation methods for each feature to address the challenge of adding noise to heterogeneous features. For numerical features, we will use the Laplace mechanism.

Laplace Mechanism: The Laplace mechanism [11] is commonly utilized in differential privacy algorithms for numerical data. Given a dataset ***D***, exist a function ***f*** that operates

on ***D***

$$f: D \rightarrow R^d, \tag{4}$$

the Laplace mechanism adds noise that is scaled to the sensitivity of Δ*f*, which is defined as:

$$\Delta f = max_{D,D':\|D-D'\|_1=1} \|f(D) - f(D')\|_1 \tag{5}$$

Randomized Response Mechanism: Assuming that the user's data comes from a categorical feature *f*$_r$, the true categorical data value of the user is denoted as v_{rk}, the

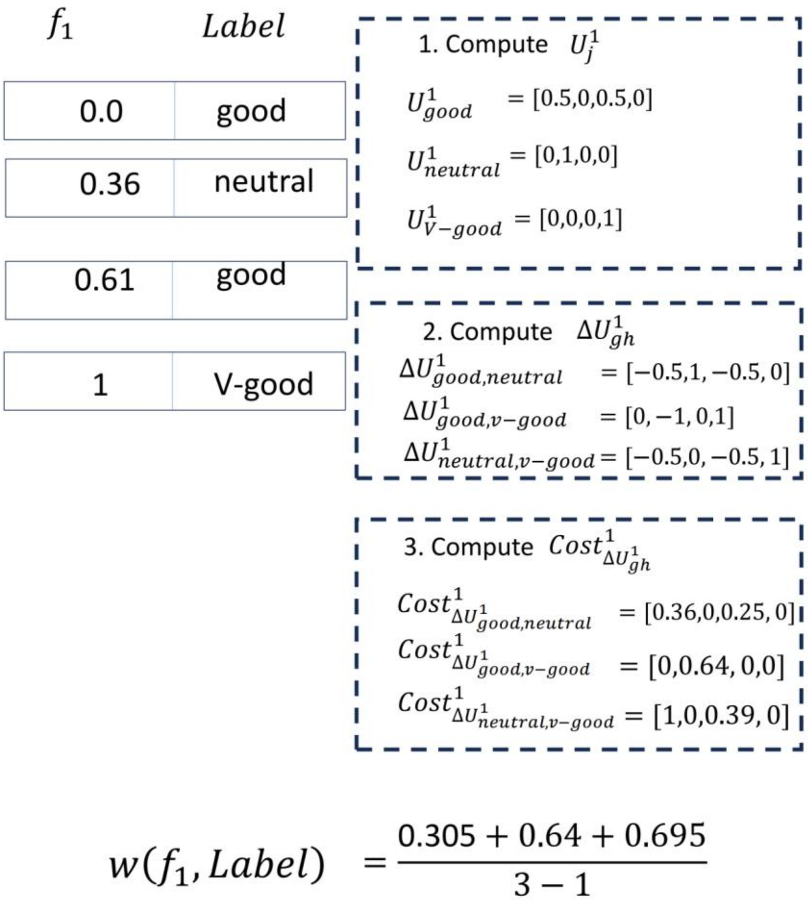
f_1
$Label$
0.0
good
0.36
neutral
0.61
good
1
V-good
1. Compute U_j^1
$U_{good}^1 = [0.5,0,0.5,0]$
$U_{neutral}^1 = [0,1,0,0]$
$U_{V-good}^1 = [0,0,0,1]$
2. Compute ΔU_{gh}^1
$\Delta U_{good,neutral}^1 = [-0.5,1,-0.5,0]$
$\Delta U_{good,v-good}^1 = [0,-1,0,1]$
$\Delta U_{neutral,v-good}^1 = [-0.5,0,-0.5,1]$
3. Compute $Cost_{\Delta U_{gh}^1}^1$
$Cost_{\Delta U_{good,neutral}^1}^1 = [0.36,0,0.25,0]$
$Cost_{\Delta U_{good,v-good}^1}^1 = [0,0.64,0,0]$
$Cost_{\Delta U_{neutral,v-good}^1}^1 = [1,0,0.39,0]$
$w(f_1, Label) = \frac{0.305 + 0.64 + 0.695}{3 - 1}$

Fig. 2. Sample dataset with only four records, the sample dataset has numerical feature *f1* and Label set with three categories.

perturbed value after applying the mechanism is denoted as g_{rk}, we define the following perturbation mechanism [45],

$$P_r = [g_{rk} = v'_{rk}] = \begin{cases} \frac{e^{\epsilon}}{e^{\epsilon}+k-1}, v'_{rk} = v_{rk} \\ \frac{1}{e^{\epsilon}+k-1}, v'_{rk} \in V_r \mid v'_{rk} \neq v_{rk}. \end{cases} \quad (6)$$

4 Experiment

We can categorize data into personal and public data based on the application domain and the entities from which data is collected. Personal data refers to data collected from specific users' devices, such as medical information and credit card records. This type of data contains sensitive information about the users. On the other hand, public data refers to data collected from social or public devices, such as emotion recognition in music, without including personal information.

Table 1. Breast Cancer Data and Different Privacy Budgets

ϵ in Random Forest Algorithm	0.2	0.4	0.6	0.8	1
Accuracy-original dataset	0.75	0.75	0.75	0.75	0.75
Accuracy-synthesis dataset	0.75	0.75	0.75	0.75	0.75
ϵ in Mixed Naive Bayes Algorithm	0.2	0.4	0.6	0.8	1
Accuracy-original dataset	0.76	0.76	0.76	0.76	0.76
Accuracy-synthesis dataset	0.76	0.76	0.76	0.76	0.76

To demonstrate the feasibility of our method, we conducted experiments on personal and public datasets. Specifically, in the public dataset, we utilized two music emotion recognition datasets: Turkish Music (TM)Emotion[9] and MER500(MER)[14]. It is worth

Table 2. Credit Approval Data and Different Privacy Budgets

ϵ in Random Forest Algorithm	0.2	0.4	0.6	0.8	1
Accuracy-original dataset	0.87	0.87	0.87	0.87	0.87
Accuracy-synthesis dataset	0.87	0.87	0.86	0.86	0.86
ϵ in Mixed Naive Bayes Algorithm	0.2	0.4	0.6	0.8	1
Accuracy-original dataset	0.80	0.80	0.80	0.80	0.80
Accuracy-synthesis dataset	0.83	0.83	0.83	0.83	0.83

Table 3. Turkish Music Emotion Data and Different Privacy Budgets

ϵ in Random Forest Algorithm	0.2	0.4	0.6	0.8	1
Accuracy-original dataset	0.73	0.73	0.73	0.73	0.73
Accuracy-synthesis dataset	0.73	0.73	0.73	0.73	0.73
ϵ in Mixed Naive Bayes Algorithm	0.2	0.4	0.6	0.8	1
Accuracy-original dataset	0.73	0.73	0.73	0.73	0.73
Accuracy-synthesis dataset	0.74	0.74	0.74	0.74	0.74

Table 4. MER500 Data and Different Privacy Budgets

ϵ in Random Forest Algorithm	0.2	0.4	0.6	0.8	1
Accuracy-original dataset	0.71	0.71	0.71	0.71	0.71
Accuracy-synthesis dataset	0.70	0.70	0.70	0.70	0.70
ϵ in Mixed Naive Bayes Algorithm	0.2	0.4	0.7	0.8	1
Accuracy-original dataset	0.65	0.65	0.65	0.65	0.65
Accuracy-synthesis dataset	0.66	0.66	0.65	0.65	0.65

Table 5. In the Mixed Naive Bayes Algorithm, comparing accuracy with the two commonly used methods within the same privacy budget.

Dataset	CA	MER	TM	BC
Accuracy- Proposed method	**0.83**	0.65	0.74	0.76
Accuracy-Information Gain	0.80	0.65	0.74	0.76
Accuracy-Chi-square	0.80	0.65	0.74	0.76

Table 6. In the Random Forest algorithm, comparing accuracy with the two commonly used methods within the same privacy budget.

Dataset	CA	MER	TM	BC
Accuracy- Proposed method	**0.86**	**0.71**	0.73	0.75
Accuracy-Information Gain	0.85	0.71	0.73	0.75
Accuracy-Chi-square	0.85	0.69	0.73	0.75

noting that music emotion data often contains subjective information during annotation, leading to lower classification accuracy when using conventional classification algorithms. Therefore, we aimed to test our method's ability to preserve the utility of the original data without compromising its effectiveness.

In the personal dataset, we employed the Breast Cancer (BC) [49] and Credit Approval (CA)dataset [19], which contain significant sensitive information. These datasets are composed of heterogeneous data, making them suitable for demonstrating the versatility of our method.

We employed two common classifiers, Random Forest [4] and Mixed Data Bayesian [18]. We used the accuracy of the classifiers as a metric to evaluate the utility of the data under different privacy budgets. By comparing the utility of the original and synthesized data, we aimed to assess the impact of different privacy budgets on the effectiveness of our proposed method. Table 1-4 shows the changes in accuracy for different datasets under different privacy budgets.

Based on the experimental results, we observed that despite injecting random noise into the original data to synthesize a new dataset, the classifier trained on the synthetic data exhibited a similar accuracy to the one trained on the original data. There was no significant decrease in utility. Additionally, by introducing random noise to features unrelated to the labels in the dataset, we improved the utility of the originally low-utility data. Compared to the two commonly used methods, Information Gain and Chi-square, our approach synthesizes data that allows the classifier to achieve higher accuracy within the same privacy budget (Table 5-6).

5 Conclusion

Today, classification data can be divided into public datasets and personal datasets. Personal datasets carry the risk of exposing sensitive information, while public datasets may have lower accuracy due to subjective annotations. This paper proposes a novel data synthesis method to address two critical challenges in clas-sification data: the threat of user information leakage in personal datasets and the difficulty of training high-performance classifiers due to insufficient training samples. By applying our method to heterogeneous data, we have found that it can handle various types of feature data. By injecting noise into features with low correlation to the labels, we can obtain synthetic datasets that have similar utility to the original data. In some low-utility original datasets, our random noise provides diversity to the samples, resulting in classifiers trained on our synthesized data achieving higher classification accuracy.Compared to the previously commonly used feature selection methods, within the same privacy budget and classifier conditions, using our proposed method to synthesize data can result in a classifier with better performance.

References

1. Abay, N.C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Sweeney, L.: Privacy preserving synthetic data release using deep learning. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18. pp. 510–526. Springer (2019)
2. Alishahi, M., Moghtadaiee, V., Navidan, H.: Add noise to remove noise: Local differential privacy for featureselection. Computers & Security **123**, 102934 (2022)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (2017)
4. Azar, A.T., Elshazly, H.I., Hassanien, A.E., Elkorany, A.M.: A random forest classifier for lymph diseases. Computer methods and programs in biomedicine **113**(2), 465–473 (2014)
5. Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Lee, R., Bhavnani, S.P., Byrd, J.B., Greene, C.S.: Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes **12**(7), e005122 (2019)
6. Bendechache, M., Kechadi, M.T.: Distributed clustering algorithm for spatial data mining. In: 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM). pp. 60–65. IEEE (2015)
7. Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M.: Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis **143**, 106839 (2020)
8. Cai, S., Zhang, Y., Luo, X., Cheung, Y.M., Jia, H., Liu, P.: Robust categorical data clustering guided by multi-granular competitive learning. In: 2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS). pp. 288–299. IEEE (2024)
9. Chen J, Ji Y, Zou R, et al. QGRL: quaternion graph representation learning for heterogeneous feature data clustering Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.: 297-306. (2024)

10. Er, M.B.: Turkish Music Emotion. UCI Machine Learning Repository (2023), DOI: https://doi.org/10.24432/C5JG93
11. Fernandes, N., McIver, A., Morgan, C.: The laplace mechanism has optimal utility for differential privacy over continuous queries. In: 2021 36th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS). pp. 1–12. IEEE (2021)
12. Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., Sales, A.P.: Generation and evaluation of synthetic patient data. BMC medical research methodology **20**(1), 1–40 (2020)
13. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans (2017)
14. Li, J., Han, L., Wang, Y., Yuan, B., Yuan, X., Yang, Y., Yan, H.: Combined angular margin and cosine margin softmax loss for music classification based on spectrograms. Neural Computing and Applications **34**(13), 10337–10353 (2022)
15. Liang, T., Wu, X., Xu, J., Feng, Q.: The distributed algorithms for the lower- bounded k-center clustering in metric space. Theoretical Computer Science **1027**, 114975 (2025)
16. Liu Y, Shang X, Zhang Y, et al. Mind the Gap: Confidence Discrepancy Can Guide Federated Semi-Supervised Learning Across Pseudo-Mismatch[J]. arXiv preprint arXiv:2503.13227, (2025)
17. Lin, Z., Khetan, A., Fanti, G., Oh, S.: Pacgan: The power of two samples in generative adversarial networks. Advances in neural information processing systems **31** (2018)
18. Murray, J.S., Dunson, D.B., Carin, L., Lucas, J.E.: Bayesian gaussian copula factor models for mixed data. Journal of the American Statistical Association **108**(502), 656–665 (2013)
19. Quinlan, J.R.: Credit Approval. UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5FS30
20. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data 6(1), 1–48 (2019)
21. Shorten, C., Khoshgoftaar, T.M., Furht, B.: Textdata augmentation for deep learning. Journal of big Data **8**, 1–34 (2021)
22. Soria-Comas, J., Domingo-Ferrer, J.: Big data privacy: challenges to privacy principles and models. Data science and engineering **1**(1), 21–28 (2016)
23. Sun, Z., Strang, K.D., Pambel, F.: Privacy and security in the big data paradigm. Journal of computer information systems (2018)
24. Van Dyk, D.A., Meng, X.L.: The art of data augmentation. Journal of Computational and Graphical Statistics **10**(1), 1–50 (2001)
25. Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., Ho, A.: Will we run out of data? an analysis of the limits of scaling datasets in machine learning. arXiv preprint arXiv:2211.04325 (2022)
26. Wang, P., Zhang, Y., Zhang, Y., Lu, Y., Li, M., Cheung, Y.m.: Clustering by learning the ordinal relationships of qualitative attribute values. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
27. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021)
28. Wu, X., Liang, L., Shi, Y., Fomel, S.: Faultseg3d: Using synthetic data sets to train an end-to-end convolutional neural network for 3d seismic fault segmentation. Geophysics **84**(3), IM35–IM45 (2019)
29. Xiaopeng Luo, Yiqun Zhang*, Yuzhu Ji, Peng Liu and Taoting Xiao, : Efficient Topology-Driven Clustering for Imbalanced Streaming Biomedical Data Analysis, Proceedings of the 2024 International Conference on Bioinformatics and Biomedicine (BIBM 2024), pp.

2262-2267, Lisbon, Portugal, December 3-6, (2024)

30. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Advances in neural information processing systems **32** (2019)
31. Yang, M., Lyu, L., Zhao, J., Zhu, T., Lam, K.Y.: Local differential privacy and its applications: A comprehensive survey. arXiv preprint arXiv:2008.03686 (2020)
32. Yoon, J., Drumright, L.N., Van Der Schaar, M.: Anonymization through data synthesis using generative adversarial networks (ads-gan). IEEE journal of biomedical and health informatics **24**(8), 2378–2388 (2020)
33. Yuzhu Ji, Mingshan Sun, Jianyang Shi, Xiaoke Jiang, Yiqun Zhang* and Haijun Zhang, “SeqPose: An End-to-End Framework to Unify Single-frame and Video-based RGB Category-Level Pose Estimation”, Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI 2025), in press, Montreal, Canada, August 16-22, (2022)
34. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: Privbayes: Private data release via bayesian networks. ACM Transactions on Database Systems (TODS) **42**(4), 1–41 (2017)
35. Zhang, Y., Cheung, Y.M.: An ordinal data clustering algorithm with automated distance learning. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). vol. 34, pp. 6869–6876 (2020)
36. Zhang, Y., Cheung, Y.M.: Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data. IEEE Transactions on Neural Networks and Learning Systems (TNNLS), Vol. 34, No. 9, pp. 6530-6544(2023)
37. Zhang, Y., Cheung, Y.m.: Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol **44**, No. 7, pp. 3560-3576(2022)
38. Zhang, Y., Cheung, Y. M., & Zeng, A. (2022, July). Het2Hom: Representation of Heterogeneous Attributes into Homogeneous Concept Spaces for Categorical-and-Numerical-Attribute Data Clustering. In IJCAI (pp. 3758-3765). (2022)
39. Zhang, Y., Zhao, M., Chen, Y., Lu, Y., Cheung, Y.m.: Learning unified distance metric for heterogeneous attribute data clustering. Expert Systems with Applications (ESWA) (2025)
40. Zhang, Yunfan, et al.: Asynchronous federated clustering with unknown number of clusters. Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 21. (2025)
41. Zhang, Y., Zou, R., Zhang, Y., Zhang, Y., Cheung, Y.m., Li, K.: Adaptive micro partition and hierarchical merging for accurate mixed data clustering. Complex & Intelligent Systems (CAIS), in press (2024)
42. Zhang, Z., Wang, T., Li, N., Honorio, J., Backes, M., He, S., Chen, J., Zhang, Y.: {PrivSyn}: Differentially private data synthesis. In: 30th USENIX Security Symposium (USENIX Security 21). pp. 929–946 (2021)
43. Zhang, Y., Member, S., Feng, S., Wang, p., Tan, Z., Luo, X.:Learning Self-Growth Maps for Fast and Accurate Imbalanced Streaming Data Clustering. IEEE Transactions on Neural Networks and Learning Systems (2025)
44. Zhao, M., Feng, S., Zhang, Y., Li, M., Lu, Y., Cheung, Y.M.: Learning order forest for qualitative-attribute data clustering. In: ECAI 2024, pp. 1943–1950. IOS Press (2024)
45. Zhao, Y., Chen, J.: A survey on differential privacy for unstructured data content. ACM Computing Surveys (CSUR) **54**(10s), 1–28 (2022)
46. Zhao, Z., Kunar, A., Birke, R., Chen, L.Y.: Ctab-gan: Effective table data synthesizing. In: Asian Conference on Machine Learning. pp. 97–112. PMLR (2021)
47. Zhu, T., Li, G., Zhou, W., Philip, S.Y.: Differentially private data publishing and analysis: A survey. IEEE Transactions on Knowledge and Data Engineering 29(8), 1619–1638

(2017)

48. Zou, R., Zhang, Y., Zhang, Y., Lu, Y., Li, M., Cheung, Y.m.: Federated clustering with unknown number of clusters. In: 2024 6th International Conference on Data- driven Optimization of Complex Systems (DOCS). pp. 671–677. IEEE (2024)
49. Zwitter, M., Soklic, M.: Breast Cancer. UCI Machine Learning Repository (1988), DOI: https://doi.org/10.24432/C51P4M