Enhancing the Performance of 3D Rotation Perturbation in Privacy Preserving Data Mining Using Correlation Based Feature Selection

Mahit Kumar Paul and Md. Rabiul Islam

Abstract A large amount of valuable data is being produced every day with the development of technologies. To retrieve knowledge and information from these data, mining and analysis are mandatory. But, the data may contain sensitive information of the individuals like medical diagnostic reports which they do not want to expose. Privacy preserving data mining, i.e., PPDM can help in this issue keeping the sensitive information private as well as preserving the data utility. Rotation-based perturbation technique contributes to satisfying both aspects of PPDM, i.e., individuals' privacy and data utility besides other PPDM techniques. In this work, we proposed a way for generating the triplet (set of three features) for 3D rotation perturbation technique using correlation among the features. This triplet generation is a fundamental step in 3D rotation perturbation technique. The analysis of information entropy, privacy protection and utility analysis elucidates that correlation-based triplet generation provides better data privacy and utility than existing triplet generation for 3D rotation perturbation technique.

Keywords Privacy · Data utility · Correlation · Perturbation · Rotation

1 Introduction

Extraction of knowledge and necessary information from data is the fundamental task in data mining, and the extracted knowledge is useful in decision-making activities [1]. But the data available for data mining tasks may contain sensitive information of the individuals, and they do not want to expose it for analysis purposes. For example, analyzing the students' semester-wise performance, we can have insight about the students' weaknesses, scope of developments, drop-out reasons and so on [2]. But, the students' detailed reports of some educational institutes are confidential and cannot be made available for data mining tasks. On the other hand, we need to analyze

M. K. Paul (✉) · Md. R. Islam
Department of Computer Science & Engineering,
Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Big Data, IoT, and Machine Learning*, Lecture Notes on Data Engineering and Communications Technologies 95, https://doi.org/10.1007/978-981-16-6636-0_17

the data to take decisions. That is why it is required to perturb the original data in some ways for analyzing. In this case, the main concerned issue is the privacy of the individuals' as well as the data utility because perturbed data may loss its underlying distribution while preserving privacy. Privacy preserving data mining (PPDM) can deal with this issue. PPDM aims to keep the data utility while maintaining the privacy of the individuals [3, 4].

PPDM targets to mitigate the risk of information leakage of the individuals and organizations. That is why, the data is released in such a way so that the intruder cannot estimate the original information, and the data utility has to be kept at the same time. Data perturbation approaches release the aggregate information regarding the data set. This aggregate information is useful to find out the knowledge through data mining algorithms. Also, this aggregate information introduces uncertainty of the individual values and thus reduces the chance of revealing private information [5].

Rotation-based perturbation approach preserves the geometric properties such as Euclidean distance and inner product of the data set [6]. Available rotation-based perturbation techniques can be of two types: two-dimensional rotation transformation (2DRT) [7] and three-dimensional rotation transformation (3DRT) [8]. In case of 2DRT, the direction of rotation is constantly orthogonal to the z-axis, i.e., the xy-plane. On the other hand, the direction of rotation can be chosen independently for 3DRT. It can be x-axis, y-axis or z-axis depending upon the inherent planes. In this work, we considered the 3DRT perturbation technique. In [8], the authors used a straightforward approach for generating the triplet for 3DRT. They simply selected three consecutive features for generating a triplet without considering the correlation among the features. We utilized the correlation among the features of the data set to generate the triplet for 3DRT in this paper.

The residual of this paper is organized in several sections. In Sect. 2, perturbation work done by many researchers in the field of PPDM is discussed. Throughout Sect. 3, the workflow of our work is described precisely. Different metrics used for privacy and data utility measurement are introduced in Sect. 4. The experimental throughput of our work is analyzed in Sect. 5. Finally, in Sect. 6, the conclusion of our work is given.

2 Related Work

Researchers have developed many methods which attempt to deal with the trade-off between preservation of privacy and maintenance of data utility in PPDM. Olivera and Zaiane proposed two-dimensional rotation-based transformation technique which is not dependent on any clustering algorithm [7]. Data features are rotated pairwise depending upon feature concerned threshold values. In [9], the authors familiarized a set of geometric data perturbation techniques such as translation, scaling and rotation-based data perturbation as well as hybridization of data perturbation. These methods only alter the sensitive features of the data set. A three-dimensional rotation perturbation technique is introduced in [8] which makes triplets of features and rotates the triplets at a time. The rotation direction can be any of the three x, y or z axes. For stream data mining, two distinguished data perturbation techniques are proposed in [10] for privacy preservation. Random projection and random translation along with two different forms of additive noise are used to develop the two perturbation techniques in [10]. Chamikara et al. proposed a non-invertible and extendable perturbation algorithm called PABIDOT for the preservation of privacy of big data. PABIDOT consists of multidimensional geometric transformations, reflection, translation and rotation succeeded by randomized expansion and random tuple shuffling [4]. For the privacy preservation of stream data and big data, $P^2 R_{\text{o}} C\text{Al}$ (privacy preserving rotation-based condensation algorithm) is proposed in [11]. $P^2 R_{\text{o}} C\text{Al}$ combines the proficiency of condensation and accuracy of rotation to deal with both of the aspects of PPDM. A new privacy preserving technique called secure and efficient data perturbation algorithm utilizing local differential privacy (SEAL) is discussed in [12]. SEAL utilizes Chebyshev interpolation and Laplacian noise. Combining these two, SEAL gives a good harmony between privacy and utility of PPDM.

3 Methodology

In this paper, a method to enhance the performance of three-dimensional rotation perturbation technique is proposed using the correlation among the features of the data set. The detailed workflow of our work is provided in Fig. 1. The input of our procedure is the normalized data set D^N. For normalization, z-score is used because it provides better performance among the other normalization techniques for the full feature set [13]. In three-dimensional rotation, the axis of rotation can be of x, y or z-axis. For three-dimensional rotation perturbation technique, double rotation matrices are used, where the data are rotated along the axis pairs xy, yz or xz and the rotation matrices are formulated such as in [8]. After selecting the axis pair, the step of triplet generation is implemented. In the existing three-dimensional perturbation technique, consecutive three features are used to generate triplets regardless of the correlation among the features. In our procedure, the correlation among the features is considered and given emphasized in the generation of triplets. The mostly correlated three features are put together to generate the first triplet, next mostly correlated three features are put together to generate the second triplet and so on. This can be explained using the correlation matrix of the BODS data set (Table 2) provided in Table 1. The correlation-based generated triplets are (Attr2, Attr3, Attr4), (Attr6, Attr7, Attr8), (Attr5, Attr9, Attr10) and (Attr2, Attr3, Attr1). In the last triplet, Attr2 and Attr3 are reused to generate triplet with Attr1. After rotation, these reused triplets are discarded. This way of triplet generation outperforms than the existing consecutive triplet generation schema which is illustrated in the performance analysis section. After triplet generation, each of the triplets is rotated for several angles θ in the range $0.1 \leq \theta < 360$. In the next step, the variances between the original and rotated features are determined for each triplet.

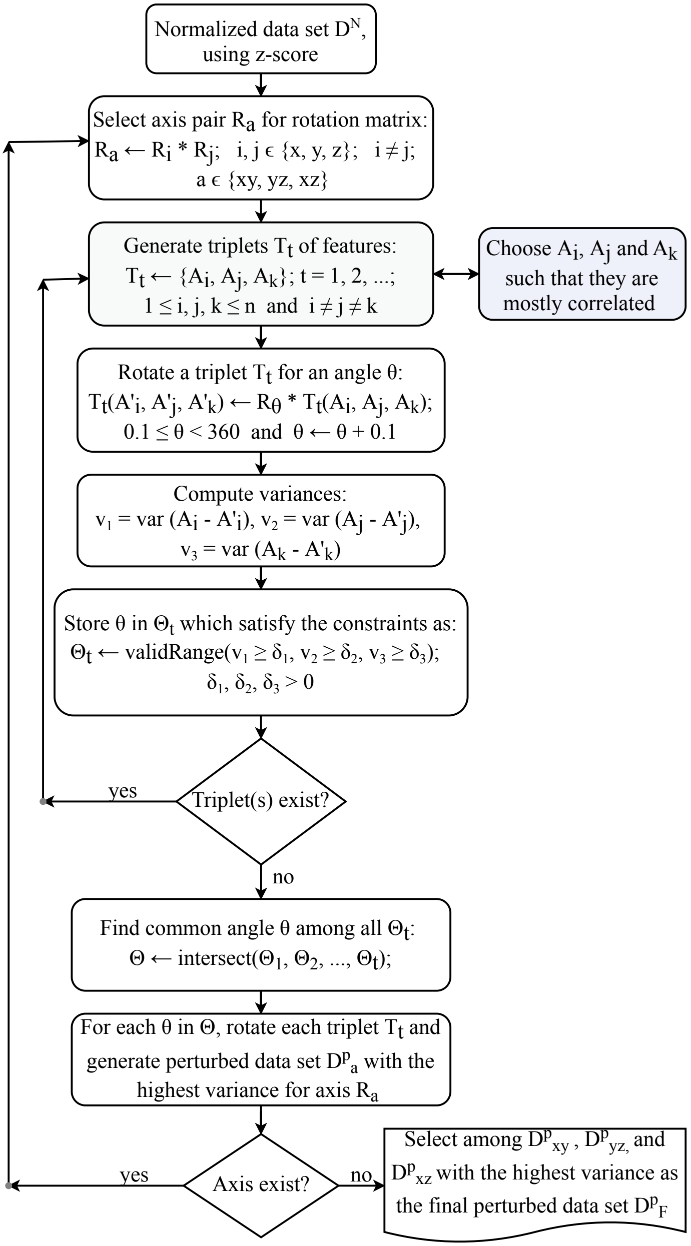
Normalized data set D^N, using z-score
Select axis pair R_a for rotation matrix: $R_a \leftarrow R_i * R_j$; $i, j \in \{x, y, z\}$; $i \neq j$; $a \in \{xy, yz, xz\}$
Generate triplets T_t of features: $T_t \leftarrow \{A_i, A_j, A_k\}$; $t = 1, 2, ...$; $1 \leq i, j, k \leq n$ and $i \neq j \neq k$
Choose A_i, A_j and A_k such that they are mostly correlated
Rotate a triplet T_t for an angle θ: $T_t(A'_i, A'_j, A'_k) \leftarrow R_\theta * T_t(A_i, A_j, A_k)$; $0.1 \leq \theta < 360$ and $\theta \leftarrow \theta + 0.1$
Compute variances: $v_1 = var(A_i - A'_i)$, $v_2 = var(A_j - A'_j)$, $v_3 = var(A_k - A'_k)$
Store θ in Θ_t which satisfy the constraints as: $\Theta_t \leftarrow validRange(v_1 \geq \delta_1, v_2 \geq \delta_2, v_3 \geq \delta_3)$; $\delta_1, \delta_2, \delta_3 > 0$
Triplet(s) exist?
yes
no
Find common angle θ among all Θ_t: $\Theta \leftarrow intersect(\Theta_1, \Theta_2, ..., \Theta_t)$;
For each θ in Θ, rotate each triplet T_t and generate perturbed data set D^p_a with the highest variance for axis R_a
Axis exist?
yes
no
Select among D^p_{xy}, $D^p_{yz,}$ and D^p_{xz} with the highest variance as the final perturbed data set D^p_F

Fig. 1 Proposed workflow

Table 1 Correlation matrix for the data set BODS

	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8	Attr9	Attr10
Attr1	1	−0.06	−0.04	−0.04	−0.07	−0.05	−0.1	−0.06	−0.05	−0.04
Attr2	−0.06	1	0.64	0.65	0.49	0.52	0.59	0.55	0.53	0.35
Attr3	−0.04	0.64	1	0.91	0.71	0.75	0.69	0.76	0.72	0.46
Attr4	−0.04	0.65	0.91	1	0.69	0.72	0.71	0.74	0.72	0.44
Attr5	−0.07	0.49	0.71	0.69	1	0.59	0.67	0.67	0.6	0.42
Attr6	−0.05	0.52	0.75	0.72	0.59	1	0.59	0.62	0.63	0.48
Attr7	−0.1	0.59	0.69	0.71	0.67	0.59	1	0.68	0.58	0.34
Attr8	−0.06	0.55	0.76	0.74	0.67	0.62	0.68	1	0.67	0.35
Attr9	−0.05	0.53	0.72	0.72	0.6	0.63	0.58	0.67	1	0.43
Attr10	−0.04	0.35	0.46	0.44	0.42	0.48	0.34	0.35	0.43	1

Then these triplet variances are compared with previously defined thresholds δ_1, δ_2 and δ_3. These thresholds help to generate a random range of rotation angles Θ_t for a specific triplet T_t. Thus, the ranges $\Theta_1, \Theta_2, \ldots, \Theta_t$ for each of the triplets $T_1, T_2, \ldots, T_t$ are generated. Then the common angle range Θ among all Θ_t is computed by intersection. For each of the angles θ in Θ, all of the triplets are rotated, and the perturbed data set D_a^P with the highest variance for the axis pair R_a is generated. In this way, the perturbed data sets D_{xy}^P, D_{yz}^P and D_{xz}^P for the axis pairs xy, yz and xz are generated. Finally, the perturbed data set with the highest variance among D_{xy}^P, D_{yz}^P and D_{xz}^P is selected and output as the final perturbed data set D_F^P.

4 Evaluation Metrics

Performance measurement of a perturbation technique is very crucial in PPDM. To asses the performance of 3D rotation perturbation technique, different privacy and data utility metrics are used in this work as follows.

4.1 Increase in Information Entropy

The information entropy of a data set can be measured by using Shannon's entropy formulation given in Eq. 1 [4].

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \tag{1}$$

where $H(X)$ is the entropy of X which is a discrete random variable with $X =< x_1, x_2, \ldots, x_n >$, $\sum$ indicates the summation over X, and $P(x_i)$ is the probability of the occurrence of x_i. The values of the average increase in information entropy are calculated by using Eq. 2 [4].

$$\mathrm{AIE} = \frac{\sum_{i=1}^{n} (H(x_i') - H(x_i))}{n} \tag{2}$$

where AIE stands for an average increase in information entropy and $H(x_i')$ and $H(x_i)$ are entropy for each feature of perturbed data set and original data set, respectively. When the values of AIE are positive, it indicates that the features of the perturbed data set contain more impurity as compared to the original data set and hence preserve more privacy.

4.2 Privacy Protection of Data

The privacy protection of data is measured using six different metrics in this paper. They are privacy [8], value difference (VD) [14], rank differences such as RP, RK, CP and CK [14]. Generally, privacy of a perturbation technique is defined as the variance between the original and perturbed data values [8]. The value difference (VD) between the original and perturbed data values is defined as Frobenius norm [14]. The rank-based privacy metric RP is used to measure the average changes of rank of all the data features [14]. RK [14] can represent the percentage of data values that preserves their individual ranks of magnitude in every feature afterward the perturbation. The metric CP is used to denote the change of rank of the average value of the features [14]. Resembling to RK, CK is used to measure the percentage of the features that preserve their ranks of the average value after the perturbation [14].

4.3 Utility of Data

To measure the utility maintenance of 3D rotation perturbation technique, three metrics are used in this paper—accuracy, $F1$-score and area under the ROC curve, i.e., AUC. Accuracy gives an insight of accurate decisions made by the data mining algorithms. $F1$-score, alternatively known as F-measure or F-score, resembles a balance between precision and recall provided by a data mining algorithm. AUC values resemble the betterment of a data object being classified between two separated groups. In ROC curve, true positive rate $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ is plotted as a function of false positive rate $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$.

5 Performance Analysis

5.1 Data Set, Classifiers and Settings

In this paper, we used five data sets with varying number of features and instances to evaluate the performance. The details of the used data set are given in Table 2. To measure the classification performance, we used C4.5 which classifies instances by generating trees. Weka is used to implement C4.5 with the default parameters. MATLAB R2020a (v9.8.0.1323502) is used as the implementation platform for three-dimensional rotation perturbation technique. We worked on a computer having the configuration as processor: Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz 1.80 GHz; RAM: 8.00 GB; system type: 64-bit operating system, x64-based processor.

5.2 Analysis of Privacy

In this paper, three-dimensional rotation perturbation technique with correlation-based feature selection is denoted as 3DRT-CF and with non-correlated, i.e., consecutive feature selection-based approach is denoted as 3DRT-NCF. In the bar graph of Fig. 2, the average increase in information entropy (AIE) values returned by 3DRT-CF and 3DRT-NCF are shown. We see that all of the AIE values are positive which indicates both 3DRT-CF and 3DRT-NCF preserve more privacy than the original data set. Alongside looking at the numeric value comparison from the bar graph in Fig. 2, it is observed that 3DRT-CF preserves more privacy than 3DRT-NCF.

Table 2 Data set used for analysis tasks

S. No.	Original data set name	Abbreviation	#Instances	#Features	Feature types
1	Electricity	ELDS	45312	8	Integer, real
2	Breast Cancer Wisconsin (original)	BODS	699	10	Integer
3	Diabetic Retinopathy Debrecen Data Set	DRDS	1151	19	Integer, real
4	Breast Cancer Wisconsin (diagnostic)	BDDS	569	31	Real
5	SPECTF Heart	SHDS	267	44	Integer

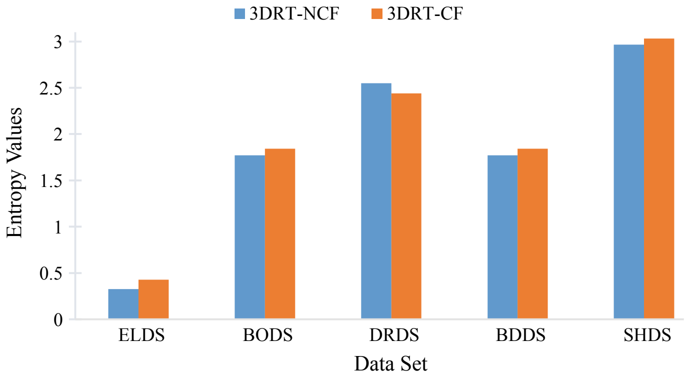

Fig. 2 Average increase in information entropy

Table 3 Privacy protection by 3DRT-NCF and 3DRT-CF

Data set	Methods	Privacy	VD	RP	RK	CP	CK
ELDS	3DRT-NCF	1275.264	1.238	17304.53	0	3.75	0
	3DRT-CF	1314.695	1.255	19337.64	0	3.5	0
BODS	3DRT-NCF	1.282	1	247.534	0.002	4	0.1
	3DRT-CF	1.361	1	244.167	0.001	3	0.1
DRDS	3DRT-NCF	255.711	1.003	437.345	0.001	5.474	0.053
	3DRT-CF	258.816	1.013	479.992	0.001	6.421	0
BDDS	3DRT-NCF	9436.237	1	234.932	0.002	11.613	0.065
	3DRT-CF	9667.235	1	255.409	0.001	13.032	0
SHDS	3DRT-NCF	1.149	1.002	114.518	0.002	14.955	0
	3DRT-CF	1.153	1.002	115.356	0.003	16.364	0

In order to show the efficacy of 3DRT-CF over 3DRT-NCF, more six privacy preservation metrics are implemented and the corresponding experimental values are provided in Table 3. The larger values of privacy, VD, RP and CP, and the smaller values of RK and CK indicate higher privacy preservation [14]. The values of privacy resemble that 3DRT-CF outperforms 3DRT-NCF for all of the data set. From the values of VD and RK, it is observed that 3DRT-CF performs better or equal with 3DRT-NCF. 3DRT-CF performs 80% better than 3DRT-NCF concerning RP values. Also, considering CP and CK values, 3DRT-CF outperforms 3DRT-NCF except for few cases.

5.3 Analysis of Utility

Alongside the preservation of privacy, the other primary property of a perturbation technique in PPDM is to keep the utility of the perturbed data set close to the original data set. To evaluate this property, we measured the accuracy, $F1$-score and AUC values corresponding to 3DRT-CF and 3DRT-NCF returned by C4.5. Figures 3, 4 and 5 represent the curves for accuracy, $F1$-score and AUC values, respectively. From Figs. 3, 4 and 5, we see that the curves for 3DRT-CF perturbed data set are more close to the the curve for original data set, and at some points the two curves lie on each other. On the other hand, the curve for 3DRT-NCF perturbed data set is far apart from the curve for original data set. Furthermore, the curves for accuracy, $F1$-score and AUC values have almost the same shape which indicate the results

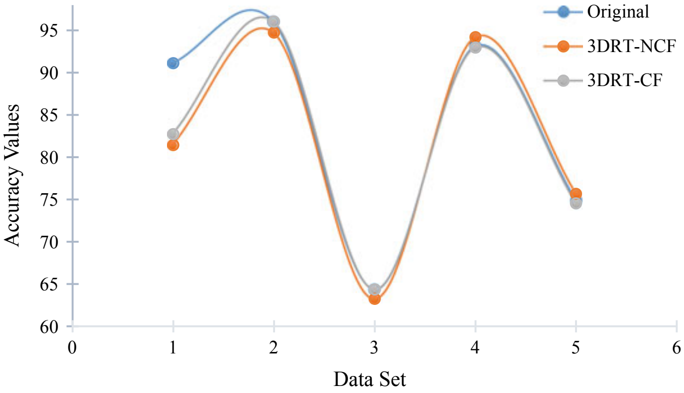

Fig. 3 Accuracy comparison

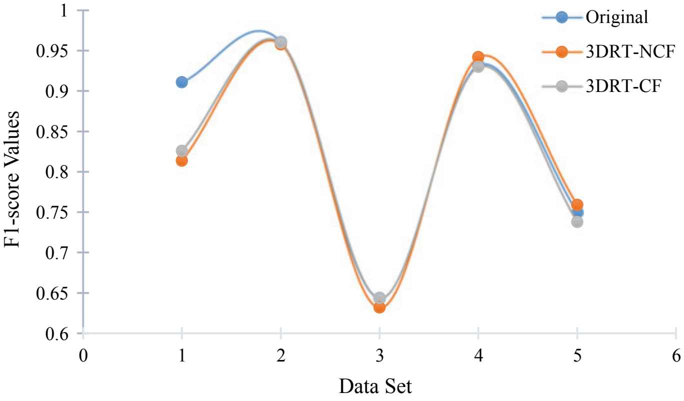

Fig. 4 $F1$-score comparison

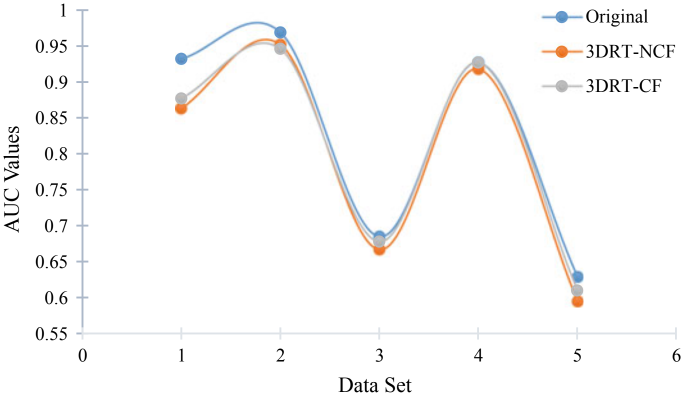

Fig. 5 AUC value comparison

are consistent. Thus, from the above analysis of privacy preservation and utility maintenance, it can be said that 3DRT-CF can perform far better than 3DRT-NCF.

6 Conclusion

In this paper, we used the correlation among the features of a data set to generate triplet (set of three features) for three-dimensional rotation perturbation technique instead of consecutive selection of features for triplet generation. Both aspects of PPDM, i.e., privacy and utility of the data, are considered while analyzing the performance. Five data sets having variations in the number of features and instances are used for experimental purposes. From the analysis of privacy and utility, it is observed that for the proposed triplet generation approach, three-dimensional rotation perturbation technique performs better than the existing triplet generation approach. Therefore, it would be a good choice to consider the correlation among the features while generating triplets. However, analyzing the effects of classifier ensembles on correlation-based triplet generation can be an extension of our work.

References

1. Salloum SA, Alshurideh M, Elnagar A, Shaalan K (2020) Mining in educational data: review and future directions. In: Advances in intelligent systems and computing. Springer International Publishing, Berlin, pp 92–102
2. Peña-Ayala A (2014) Educational data mining: a survey and a data mining-based analysis of recent works. Expert Syst Appl 41:1432–1462
3. Afrin A, Paul MK, Sattar AHMS (2019) Privacy preserving data mining using non-negative matrix factorization and singular value decomposition. In: 2019 4th International conference on electrical information and communication technology (EICT). IEEE, pp 1–6

4. Chamikara MAP, Bertok P, Liu D, Camtepe S, Khalil I (2020) Efficient privacy preservation of big data for accurate data mining. Inf Sci 527:420–443
5. Agrawal R, Srikant R (2000) Privacy-preserving data mining. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data—SIGMOD’00. ACM Press, pp 439–450
6. Chen K, Liu L (2010) Geometric data perturbation for privacy preserving outsourced data mining. Knowl Inf Syst 29:657–695
7. Oliveira S, Zaiane O (2004) Data perturbation by rotation for privacy-preserving clustering. University of Alberta Libraries
8. Upadhyay S, Sharma C, Sharma P, Bharadwaj P, Seeja KR (2018) Privacy preserving data mining with 3-D rotation transformation. J King Saud Univ Comput Inf Sci 30:524–530
9. Stanley RMO, Osmar RZ (2010) Privacy preserving clustering by data transformation. J Inf Data Manage 1:37
10. Denham B, Pears R, Naeem MA (2020) Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining. Expert Syst Appl 152:113380
11. Chamikara MAP, Bertok P, Liu D, Camtepe S, Khalil I (2018) Efficient data perturbation for privacy preserving and accurate data stream mining. Pervasive Mob Comput 48:1–19
12. Chamikara MAP, Bertok P, Liu D, Camtepe S, Khalil I (2019) An efficient and scalable privacy preserving algorithm for big data and data streams. Comput Secur 87:101570
13. Singh D, Singh B (2020) Investigating the impact of data normalization on classification performance. Appl Soft Comput 97:105524
14. Xu S, Zhang J, Han D, Wang J (2006) Singular value decomposition based data distortion strategy for privacy protection. Knowl Inf Syst 10:383–397