UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Privacy Preserving Data Mining Based on Geometrical Data Transformation Method (GDTM) and K-Means Clustering Algorithm

Nur Athirah Jamadi, Maheyzah Md Siraj, Mazura Mat Din, Hazinah Kutty Mammy, Norafida Ithnin
Information Assurance and Security Research Group (IASRG),
Faculty of Computing,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
nurathirahjmd91@gmail.com, maheyzah@utm.my, mazura@utm.my, hazinah@utm.my, afida@utm.my

Submitted: 15/11/2017. Revised edition: 16/05/2018. Accepted: 21/05/2018. Published online: 31/05/2018

***Abstract*—In current era of sharing unlimited digital information via the network, protecting the privacy of information is crucial even during the data mining process due to a high possibility of the information security risks such as being abused or leakage. Such problems motivate the research in Privacy Preserving Data Mining (PPDM) and it became one of the newest trends. Therefore, this papers reviews the related works in terms of issues, approaches, techniques, performance quantification as well as thorough discussions on pros and cons of previous researches. We also propose an improved PPDM that applying Geometrical Data Transformation Method (GDTM) and K-Means Clustering Algorithm for optimum accuracy of mining and zero data loss while preserving the privacy of information.**

***Keywords*—Privacy Preserving Data Mining, PPDM, GTDM, K-Means, taxonomy**

I. INTRODUCTION

Data Mining is an emerging research field that extracts valuable information from data that is collected and stored in large joint databases. These databases are assumed to be accessible to third party. In addition to that, during the data analysis, various data mining algorithm are applied for knowledge discovery purposes. This might lead into learning sensitive data too. Due to these, privacy breach over sensitive information becomes the main concern. Privacy preserving is a vital complementary technique, applied to data mining process. The term privacy preserving data mining (PPDM) was introduced as a solution to privacy disclosure. The aim of PPDM is to extract the useful knowledge from large dataset while protecting the sensitive information simultaneously. It deals with protecting the privacy of sensitive data without sacrificing the data utility.

There are three levels of privacy preserving data mining. The first level is the protection of sensitive data during the data formatting such as name, id number, address and others for the analytical purposes. The second level is data sanitization through various processes before revealing it to unauthorized third party. The third level is the preserving data mining that is known as knowledge hiding which protects the sensitive knowledge produced by data mining analysis [1, 2]. Main concern is on how to effectively hide sensitive values in dataset while remaining the data usability. As stated in [2], there are two broad categories of privacy preserving data mining which are first, approaches that protect the sensitive data during the mining process and second, and approaches that protect the sensitive data mining results after applying data mining algorithm.

II. ISSUES AND CHALLENGES IN PRIVACY PRESERVING DATA MINING

Taxonomy of PPDM is presented in the Fig. 1. Based on the figure, there are three main categories that lie under the term privacy preserving data mining which are privacy preserving, data mining and data distribution. In the privacy preserving category, various approach has been proposed which are perturbation, anonymization, randomized response, condensation and also cryptography. Different issues and challenges is inherited by applying privacy preserving such as privacy, homogeneity attack, data utility, information loss and high computational resource. While in the second category, data mining can be classified as clustering, classification and association rules. There are many techniques and algorithms introduced in data mining; however, they elicit several issues and challenges that should be overcome such as accuracy, scalability and efficiency. In addition, there are two categories of data distribution which are centralized and distributed.

The main idea of data mining is to extract and learn new knowledge from a large database. However, the results of data mining analysis may produce a complete or accurate data which consists of sensitive data, due to this, privacy issue arise. To ensure that data mining results remain confidential, privacy preserving technique should be applied [3]. Privacy preserving technique is important to be applied in data mining process to ensure the sensitive data remain hidden and protected from any risk of disclosure by an unauthorized party.

In order to preserve the privacy of the data, it should be modified or perturbed to an extent that, it is dissimilar and unidentifiable from the original data. This could be achieved by applying the existing privacy preserving techniques. As illustrated in Figure 1.0, there are several major techniques proposed by different researchers which are perturbation, anonymization, randomized response, condensation and cryptography. Although there are many existing techniques in preserving privacy of a dataset, it still does not provide assurance over the trade-off issues and challenges aroused.

Perturbation technique can be described as modifying or changing the data in a specific way to disguise and masquerade the sensitive data values. The original data values are replaced with synthetic values such that even after applying perturbation, the statistical computational between the original and perturbed data remained with the least significant changes [2]. Thus, it does not correspond to real-world data in return preventing from revealing the sensitive data values while simultaneously preserve the privacy of the essential data required for data analysis. However, the use of this technique involved a trade-off between privacy and data utility [4].

Anonymization refers to an approach of hiding or removing the identity of data owner. However, the sensitive data is remained for analysis. Specifically, k-anonymity model which uses generalization and suppression is proposed. Generalization replaces the associated value of data with less specific but semantically consistent value while suppression involves blocking the values. Despite this, anonymization suffers in information loss. The data is irretrievable once it is anonymized [2]. Besides, k-anonymity is prone to homogeneity attack and background knowledge attack [2, 5, 6]. In addition, anonymization cannot be applied to high-dimensional data because it will present an information loss and it required a special method to be able to be published more than once [5].

Randomized response technique is a simple technique which can be easily implemented which is done by scrambling data in such a way that the central cannot tell with probabilities better than a pre-defined threshold, whether the data contains truthful information or false information [2, 6]. This technique is based on aggregate values from the scrambled data which can be estimated with a good accuracy. However, the limitation of this is that, it treats all the records equal despite their local density. This inevitably leads to adversarial susceptible attacks on outliers' records. To prevent this, noise addition should be added but it will reduce the data utility and produces high information loss thus degrading the performance of database [7].

Condensation approach constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters [2, 6]. This approach is suitable in dealing with classification problem as the use of pseudo-data provides an additional layer of protection because it is harder to perform attack on a synthetic data. It provides improved privacy preservation prior to using pseudo-data. Main drawback of this technique is huge information loss due to the condensation of a larger number of records into a single statistical group entity

Cryptographic technique deals with avoiding sensitive information disclosure that are meant for multi-party collaboration in computing results or share sensitive data mining results [2]. Cryptographic technique offers a well-defined privacy model upon data mining. It provides privacy preserving in an in-depth measure compared to other techniques. However, cryptographic technique involves a high computational cost and resources and it does not protect the results of the computational.

In addition to the issues mentioned in privacy preserving techniques, data mining analysis also often present several issues such as illustrated in Figure 1.0; accuracy, scalability and efficiency. Accuracy is a major issue in data mining analysis. It is also closely related to information loss. The lesser loss of information, the higher the data utility produced [8]. Once the real-world dataset is collected, the application of data mining might cause errors due to inaccurate measurement or missing values [9]. Thus, pre-processing is often required before head. Besides, the selection on specific data mining task and algorithm application is also important. As the existing data mining task is designed and proposed to deal with specific requirements and parameters.

The issues and challenges in data mining field also involve scalability and efficiency. These two issues are often related to the size of the dataset used. A large data set is often used in data mining analysis; thus, the application of data mining task should be able to perform it task upon the dataset used regardless of the size [10]. There are some data mining algorithm that are not scalable and efficient such as the EM clustering algorithm.

In coherent to the discussed issues and challenges in PPDM, this project is proposed in aiming to tackle two major issues which are privacy and accuracy because in PPDM, these two issues serve as a trade-off upon each other. There are several privacy-preserving techniques that offer privacy but they do not ensure accuracy of the processed data.

Thus, the research issues here are:

1) Improve privacy of dataset used to protect the sensitive information associated before performing data mining analysis.

2) Ensure the accuracy performance of data mining analysis results.

A. Improving the Privacy of Dataset Used

This research issue has been implemented in the data privacy preserving process before conducting the data mining process. Generally, the privacy preserving technique applied will protect the sensitive attributes in a dataset. Various techniques have been proposed to deal with this issue, based on a study by [11], the techniques introduced in privacy preserving such as anonymity, random perturbation, randomized response, condensation approach and cryptography technique help in improving data privacy but each of them associated with weaknesses that could affect the overall performance of data mining process. However, these techniques can be combined to achieve a better privacy coverage and performance accuracy.

B. Ensuring the Accuracy Performance of Data Mining Analysis Results

This research issue has been addressed in data mining process after applying the privacy preserving technique. Main concern in data mining analysis is to get accurate results. However, when privacy preserving is applied, it may affect the overall data statistical result. In achieving better accuracy, a proposed technique [12, 13] has been implemented as the base of privacy preserving and verified with data mining algorithm process. Researchers in [12] come out with a data perturbation which applied geometrical transformation technique for privacy preservation and K-means clustering algorithm for data mining process. While on the other hand, work in [13] proposed a hybrid approach of privacy preserving of data transformation and K-means clustering for the verification of data mining accuracy performance checking. Among clustering algorithm, K-means is the most used in data mining analysis [9, 10].

III. Privacy Preserving Data Mining Techniques

Generally, there are five major categories of privacy preserving techniques i.e. anonymity, perturbation, randomized response, condensation and cryptographic techniques. Based on Table I, an overview of the techniques and weaknesses are detailed. Despite the general weaknesses appointed, there are lots of techniques that can be used to enhance the process of privacy preserving. Based on the privacy preservation techniques used, data mining process can be applied as a proceeding process.

In improving privacy preserving, there are a lot of enhancements techniques that can be applied to improve the results of privacy preservation technique. An efficient improvement of masking technique has been proposed [3].

TABLE I. Privacy Preserving Techniques And Weaknesses

Types	**Weaknesses**
Anonymity: This technique removes the sensitive attributes of identifier in dataset to ensure the anonymity of the record owner.	This technique does not provide full coverage on attributes disclosure. There are two types of attacks can be conducted against this technique which are the homogeneity attack and background knowledge attack [6].
Perturbation: This technique applies data pre-processing by replacing the data values with synthetic data that does not resembles any real-world data. This makes the data meaningless to human observant but only valid for statistical analysis purposes [2].	This technique despite preserving data privacy, it does not provide data utility as it does not able to reconstruct the original data values but only the distribution [11].
Randomized Response: This technique scrambled data in such a way so that the attacker cannot differentiate with a predetermined threshold probability whether the data contains true or false information.	This technique treats all the records equally irrespective of their local density. Due to this, the outlier record is susceptible to adversarial attack compared to the higher dense data regions [2].
Condensation: This technique constructs a group of clusters and then generate a pseudo-data based from the clusters statistics. Furthermore, a synthetic data is generated using pseudo-data with the same aggregate distribution of the original data.	Although this technique makes it hard for attacker to perform adversarial attacks on the synthetic data, it indirectly affects the data mining results due of information loss. This information loss happens as because condensation compressed a larger number of records into a single statistical group entity [2].
Cryptographic: This technique is applied in dealing with data sharing with multi-party scenario where they collaborate to compute results or share non-sensitive data mining results and to avoid any data disclosure. It is a protocol designed based on encryption known as Secure Multi-parties Computational (SMC).	Despite the encryption protocol employed in privacy preserving, attacker may be able to penetrate the protocol layer and breach the data confidentiality. Besides, it does not provide data usability and high computational overhead [11].

Due to the privacy concern over data mining results, they proposed two perturbative masking techniques namely data transformation and bit transformation technique in protecting sensitive numeric values in a large dataset. Firstly, they select the sensitive attribute based on the dataset. Secondly, they apply the sensitive data modification techniques proposed and measure in terms of the statistical properties, accuracy of privacy protection and clustering accuracy. Lastly, they apply the modified dataset with K-means clustering algorithm for further data mining accuracy performance measure.

Another research has been conducted by [11], they applied a hybrid approach of privacy preserving by using two combinations of privacy preservation techniques which are data randomization and data generalization on a Health Insurance Portability and Accountability Act (HIPAA) dataset. They consider a two-fold way approach by first applying the data randomization on the original data, and then apply data generalization on the randomized or modified data. The techniques can protect the data privacy, reconstruct the original data with no information loss and remain data usability.

An investigation upon data distortion is conducted by [14]. The main objective of this approach is to distort the data values while preserving the performance of data mining technique even after data distortion. In this approach, the truncated non-negative matrix factorization (NMF) with sparseness constraints is applied for data perturbation. The experiment deals with negative values. For further accuracy performance, K-nearest neighborhood (KNN) is applied.

A random data perturbation technique is proposed to preserve privacy [15]. This perturbation is conducted by adding random noise to the datasets to ensure that despite the noise addition, the originality of the data could still be preserved and accurately estimated. The process includes adding Gaussian noise and filtering method which manage to distinguish between eigenvalues noise and the corresponding original data eigenvalues. This was stated to be able to reconstruct the distribution of the actual data.

There are also several techniques, applied in enhancing the privacy preservation technique related to data perturbation by conducting geometrical data transformations. In research done by [16] they introduced a new technique known as Geometrical Data Transformation Methods (GTDMs). In their work, they provide an experiment in comparing the results of geometrical data transformations with benchmark technique; additive perturbation and with the results of K-means clustering analysis without applying transformation. The methods used is claimed to be able to balance the trade-off between privacy and accuracy.

Accordingly, in 2007, another work was done employing the technique introduced by [16] in [17]. A hybrid data transformation is applied in addressing privacy breaching by unauthorized secondary use of information. Although [16] work only tackle numerical data, they focus on categorical data instead. In order to deal with categorical data, due to the limitation of the technique, they convert the categorical data into binary data and later were transformed using GTDMs. Further, K-means clustering is performed upon the transformed data.

In coherent to this, [18] further proposed a hybrid approach of data transformation between Data Transformation Translation and Rotation (DTTR) and Data Transformation Scaling and Rotation (DTSR). Like [17], a conversion from categorical data to binary data is performed. The steps include rotation, scaling, translation and adding noise. The measures taken into consideration are effectiveness, privacy, identification of sensitive attributes and accuracy of clusters group.

A work focusing on privacy preserving in K-means clustering by clustering rotation is proposed by [19]. Based on this, data perturbation is employed. The dataset is clustered by K-means clustering. The clusters result was later perturbed by performing geometric transformations on each cluster. This transformation is particularly achieved by applying cluster displacement and rotation. In this case, the performance of privacy, data hiding, misclassification error, data usability is measured. It is proven that this specific approach is hard to reverse-engineered but still is relevant for cluster analysis.

In accordance to work by [13], they also proposed a hybrid privacy preserving technique by using data transformation method. The privacy preserving clustering was conducted on a centralized data distribution. The methods applied are Double-Reflecting data perturbation (DRDP) and Rotation Based Translation (RBT). The experiment was aimed to preserve secrecy of the dataset attributes without trading off accuracy. However, it was limited to numerical attributes only [13, 16]. As an evaluation basis, K-means is applied to check for the overall data quality.

Another approach by [20] tackles the issue of privacy preserving in multi-partitioned datasets by implementing data perturbation technique. In this research, they conducted data perturbation with corroboration and substantiation. Specifically, the proposed methods elevate the uncertainty ambiguity among horizontally and vertically partitioned data distribution. Thus, producing reliable results in terms of data privacy and transparency, adversary rate and clustering efficiency. It is also claimed that the most used data perturbation technique is Gaussian distribution. Similarly, [21] proposed a reconstruction based of data perturbation in a multi-dimensional data. The experiment was conducted using a two-dimensional dataset. The process includes, applying data perturbation algorithm to the original data. Then apply clustering algorithms which are K-means with the K-Nearest Neighbor (KNN) into the perturbed data.

In dealing with privacy preserving data mining, [22] proposed data obfuscation scheme in data statistics and data mining to gain balance between data privacy and usability. Data obfuscation is a simple version of encryption scheme but with lower security and higher usability. According to them, an accurate data noise is added to obfuscate the data. This accurate noise is generated by an improved cloud model. The method proposed succeeded in showing that the data is obfuscated correctly, efficiently and secretly. However, major limitations of this approach include data is order sensitive and requires a large number of keys.

Another work by [23] is proposed to improved method using personalized generalization which associated with anonymization technique, particularly k-anonymity. This includes transformation of quasi-identifier, applies fuzzy-based transformation on the dataset and personalizing the sensitive attribute values. The result shows a better privacy measure and reduces information loss.

Recent research proposed a simple Principal Analysis Component (PCA) in preserving privacy and accuracy of data mining analysis before and after application based on data transformation approach. The accuracy evaluation clustering analysis is based on K-means algorithm. The proposed method is compared to the traditional data transformations and is able to surpass in providing better performance in terms of inforSmation hiding and accuracy analysis [24, 25].

IV. Justification on Selected Approach in PPDM

In this section, a detailed explanation on justification upon the selection of privacy preserving data mining approach will be presented. Based on Table I presented in the Section III, existing techniques of privacy preserving does not provide a whole package of privacy in a comprehensive way. There is always a trade-off among different measures. Thus, this paper is aimed to achieve privacy and accuracy of the data by implementing a privacy preserving technique of Geometrical Data Transformation Method (GDTM) and K-means clustering algorithm.

A. GDTM as Privacy Preservation Technique

Generally, the GDTM technique is introduced by [16]. GDTM was introduced to address the privacy concern over clustering data mining analysis. This particular approach was designed to ensure privacy by data transformations could be applied without actually contributes to the loss of data mining accuracy and efficiency particularly in the used of clustering data mining task analysis.

Main problem in applying data perturbation is, once the data attributes is distorted, by further applying data mining will provide a major differences from the original dataset. Thus, this removes the accuracy validation upon the result from the data mining analysis. To provide a safeguard upon this problem, a uniform distribution should be applied to all data attributes. This could be addressed by applying GDTM.

In GDTM, there are three major functions which are Translation, Scaling and Rotation. The combination of these functions offers a higher privacy preservation technique where each one of the functions complements each functions weakness. Particularly, it minimizes the data from privacy loss and information loss [26].

Additionally, the translation and rotation function preserve the distance relationship. The core of GDTM is it will be a perturbation-invariant once the distance is preserved. In some cases, the occurrences of a distance-interference attack might happen [4]. Thus, with the translation and rotation function, the distance can be preserved approximately indirectly increases the resilience towards distance-interference attacks.

B. The Implementation of K-Means Clustering

In this section will provide a description of the selected partitioned-based clustering which is K-means clustering. In general, K-means algorithm was pioneered by an author named J. B. MacQueen [1]. It is an unsupervised learning algorithm which is mostly used in data mining and pattern recognition. It is also known as one of the simplest data mining clustering algorithms that employed the Euclidean distance function. In fact, it has advantages of briefness, efficiency and celerity.

K-means group, a provided dataset into several clusters, in the name notation itself the '*K*' letter referred to how many clusters are required in the process. In K-means clustering process, it requires a predefined initialization of *K* clusters and a central centroid for each cluster [19, 27]. The implementation of K-means clustering is aimed to provide an accuracy measure of the perturbed data once the GDTM technique is applied to the attacks.

C. Min-Max Normalization

Additionally, to achieve a better accuracy distribution of the data, a min max normalization function is further applied. Min max normalization preserves the relationship among the original data values [12, 28]. In this section, it introduces the normalization function which is applied after the privacy preserving method of GDTM is performed. Generally, the min-max normalization performs a linear transformation on the original data. In order to map a value, *v* of an attribute *A* from range [min_A, max_A] to a new range which are [new_minA, new_maxA], thus the computation is given as Equation 1.

$$\frac{v - min_A}{min_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new_{min_A} \tag{1}$$

V. The Dataset

The dataset used for this research is Diabetes Readmission Predictions provided by UCI Repository of Machine Learning. This dataset is a collection of various hospitals in the duration of 10 years data collection.

VI. Standard Performance Measurement In PPDM

There are several performance measurements that are evaluated in the basis of conducting privacy preserving data mining approach which is privacy, accuracy, efficiency, data loss, data utility, etc. It is important to implement an effective privacy preserving data mining technique which can deal with an efficient method of removing undesired challenges towards achieving the good performance of the approach. In this project, the performance measure is based on data privacy and accuracy. In order to get the measurement, a performance metric will be used for evaluation procedure. The metrics is presented in the following subsections.

A. Privacy

Privacy is a main concern in data mining. It is the measure for indicating how closely the original value of an attribute can be estimated or predicted [2]. However, the higher the privacy level, the higher the data loss produced. In order to achieve a desired privacy level, the privacy level and the data loss should be balanced.

In order to measure privacy level, there are specific metric used. Generally, the quantification used to measure data privacy is the degree of uncertainty in accordance to which original private data can be estimated. The higher the degree of uncertainty is achieved, the higher the protection of data privacy [8]. As in this project particularly, privacy preservation by data perturbation technique is measured based on the variance between the original data and the perturbed data values. This measure is presented by *Var(X-Y)* where *X* represents a single original attribute and *Y* is the distorted attribute. In accordance to this, the level of privacy can be measured by the following metric.

$$Sec = \frac{Var(X-Y)}{Var(X)} \qquad (2)$$

The value of *Sec* can be computed using the following function.

$$Sec = Var(|X_i - Y_i|)/Var(|Y_i|) \qquad (3)$$

Where *X* represents the original data, while *Y* represents the perturbed data. This *Sec* function is indicated as the degree of perturbation. Thus, *Sec* can be computed as to show the privacy measure of the proposed technique [16, 19].

B. *Accuracy*

Accuracy is often closely related to data loss which often come as a trade-off results of applying privacy preserving technique. Hypothetically, the lesser is the data loss, the better is the data accuracy [8]. Besides, accuracy measures are bounded with data loss. Data loss measure depends on the original distributions and the reconstructed distributions [13].

In this research, particularly a performance metric of Misclustering Error (M_E) percentage is used. This performance metric is measured based on the obtained perturbed data aligned with the clustering analysis due to the respective proposed technique. Misclassification error is measured in terms of the percentage of legitimate data points that are not well-satisfied in a distorted database. To get a best performance, the misclustering error should be 0%. The misclustering error, denoted by M_E is measured by,

$$M_E = \frac{1}{K} \times \sum_{i=1}^{K}(|C_i| - |C_i'|) \qquad (4)$$

Where *K* is the number of clusters in the dataset, and C_i is the unperturbed cluster while C'_i refer to the perturbed cluster [16, 19]. It is necessarily to consider both number of data points in each cluster and the actual cluster of each point. The comparison should be performed based on before and after application of data distortion.

VII. Discussion

The privacy preserving in data mining is implemented in order to prevent and protect the confidential data from unauthorized party and secondary use of information. However, privacy preservation is not easily achieved just by implementing the desired technique. Each of the existing techniques is introduced to deal with different problems from different perspectives. One technique may be able to provide privacy but may not be efficient ones. As privacy preserving technique needs to deal with issues such as information loss, ineffectiveness, inaccuracy, low privacy etc. Due to this, in order to improvise the privacy preserving technique, the related issues need to be addressed. Basically, privacy preserving techniques can be categorized into five which are anonymization, perturbation, randomized response, condensation and cryptographic technique. In this project, the focus will be on the data perturbation technique employing geometrical data transformation methods and K-means clustering data mining algorithm. This technique and algorithm is selected together because it has many advantages in order to overcome the issues and accomplish the desired goals. Besides, this method also has been proposed and applied in some of recent research in tackling the privacy and accuracy issue lays under privacy preserving data mining [12, 16, 17, 19]. In brief, the advantages of these proposed techniques are

- *Complement*: Geometrical data perturbation methods although not comprehensive, but it complements the previous related research on data perturbation. Specifically enhance the use of data perturbation technique [16].
- *Capability*: This method is capable of preserving the relationship between original data and perturbed data by applying geometrical data transformation methods with min-max normalization [12].
- *Accurateness*: This method was proved in 0% misclassification error which proved that the technique proposed is capable of handling error [19].

Based from the literature review and the brief advantages stated of the proposed technique of geometrical data transformation methods and K-means clustering, it will assist in producing an improvised empirical experiment analysis in dealing with the issues and challenges related to the privacy preserving data mining thus, accomplishing the project goals.

VIII. Conclusion

This paper presents a literature review on privacy preserving data mining, category of clustering and highlights some issues and challenges of the existing techniques. In addition, it includes an insight on the requirements that shall be applied in the experimental phase.

References

[1] Li, Y., & Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, 25, 1104-1109.

[2] Malik, M. B., Ghazi, M. A., & Ali, R. (2012). Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects, 26-32. doi: 10.1109/iccct.2012.15.

[3] Vijayarani, S., & Tamilarasi, A. (2011). An Efficient Masking Technique for Sensitive Data Protection. *International Conference on the Recent Trends in Information Technology (ICRTIT), 2011.*

[4] Patel, A., Dodiya, K., & Pate, S. (2013). A Survey on Geometric Data Perturbation in Multiplicative Data Perturbation. *International Journal of Research in Advent Technology*, 1(5), 603-607.

[5] Singh, A. P., & Parihar, M. D. (2013). A Review of Privacy Preserving Data Publishing Technique. International Journal of Emerging Research in Management &Technology.

[6] Wang, J., Luo, Y., Zhao, Y., & Le, J. (2009). A Survey on Privacy Preserving Data Mining. 111-114. doi: 10.1109/dbta.2009.147.

[7] Borhade, S. S., & Shinde, B. B. (2014). Privacy Preserving Data Mining Using Association Rule With Condensation Approach. *International Journal of Computer Science & Information Technologies,* 5(2).

[8] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in Privacy Preserving Data Mining. *ACM Sigmod Record*, 33(1), 50-57.

[9] Dalal, M., & Harale, N. (2011). A Survey on Clustering in Data Mining. Paper Presented at the Proceedings of the International Conference & Workshop on Emerging Trends in Technology.

[10] Jain, A. K., & Maheswari, S. (2012). Survey of Recent Clustering Techniques in Data Mining. *Int. J. Comput. Sci. Manage. Res,* 1, 72-78.

[11] Lohiya, S., & Ragha, L. (2012). Privacy Preserving in Data Mining Using Hybrid Approach. 743-746. doi: 10.1109/cicn.2012.166.

[12] Manikandan, G., Sairam, N., Saranya, C., & Jayashree, S. (2013). A Hybrid Privacy Preserving Approach in Data Mining. *Middle-East Journal of Scientific Research*, 15(4), 581-585.

[13] Li, L., & Zhang, Q. (2009). A Privacy Preserving Clustering Technique Using Hybrid Data Transformation Method. Paper Presented at the Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on.

[14] Kabir, S. M., Youssef, A. M., & Elhakeem, A. K. (2007). On Data Distortion for Privacy Preserving Data Mining. Paper presented at the Electrical and Computer Engineering, 2007. *CCECE 2007. Canadian Conference on*.

[15] Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2005). Random-data Perturbation Techniques and Privacy-preserving Data Mining. *Knowledge and Information Systems,* 7(4), 387-414.

[16] Oliveira, S. R., & Zaiane, O. R. (2010). Privacy Preserving Clustering by Data Transformation. *Journal of Information and Data Management,* 1(1), 37.

[17] Natarajan, A., Rajalaxmi, R., Uma, N., & Kirubhakar, G. (2007). A Hybrid Data Transformation Approach for Privacy Preserving Clustering of Categorical Data Innovations and Advanced Techniques in Computer and Information Sciences and Engineering (pp. 403-408): Springer.

[18] Rajalaxmi, R., & Natarajan, A. (2008). An Effective Data Transformation Approach for Privacy Preserving Clustering. *Journal of Computer Science,* 4(4), 320.

[19] Dhiraj, S. S., Khan, A., Khan, W., & Challagalla, A. (2009). Privacy Preservation in k-means Clustering by Cluster Rotation. Paper Presented at the TENCON 2009-2009 IEEE Region 10 Conference.

[20] Prakash, V., Shanmugam, A., & Murugesan, P. (2012). Efficient Cluster Based Privacy Preservation Data Perturbation Technique in Multi-Partitioned Datasets. *European Journal of Scientific Research,* 86(2), 254-263.

[21] Tiwari, P., & Gupta, H. (2012). Reconstruction of Perturbed Data USing K-Means. *Journal of Global Research in Computer Science,* 3(10), 18-21.

[22] Yang, P., Gui, X., Tian, F., Yao, J., & Lin, J. (2013). A Privacy-Preserving Data Obfuscation Scheme Used in Data Statistics and Data Mining. Paper presented at the High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on.

[23] Poovammal, E., & Ponnavaikko, M. (2009). An Improved Method for Privacy Preserving Data Mining. Paper Presented at the Advance Computing Conference, 2009. IACC 2009. IEEE International.

[24] Byun, J.-W., Kamra, A., Bertino, E., & Li, N. (2007). Efficient k-anonymization using Clustering Techniques Advances in Databases: Concepts, Systems and Applications (pp. 188-200): Springer.

[25] Divecha, H., & Mehta, S. (2014). Privacy Preserving Based on Geometric Transformation Using Data Perturbation.

[26] Keyvanpour, M., & Moradi, S. S. (2011). Classification and Evaluation the Privacy Preserving Data Mining Techniques By Using A Data Modification-based Framework. arXiv preprint arXiv:1105.1945.

[27] Chadha, A., & Kumar, S. (2014). An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K. Paper Presented at the Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on.

[28] Jain, Y. K., & Bhandare, S. K. (2011). Min Max Normalization Based Data Perturbation Method for Privacy Protection. *International Journal of Computer & Communication Technology (IJCCT),* 2(8), 45-50.