Turkish Journal of Computer and Mathematics Education Vol.12 No.11 (2021), 5510-5517
Research Article

An Efficient Probabilistic Multi Labeled Big Data Clustering Model for Privacy Preservation Using Linked Weight Optimization Model

Sreenivasulu Bolla[a], R. Anandan[b]

[a]Research Scholar, Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India.
Assistant Professor, Department of Computer Science and Engineering, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India.
[b]Professor, Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies, Chennai, Tamil Nadu, India.

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: An unsupervised analysis of the classification and clustering of data is one of the most powerful and insightful data mining approaches used in different disciplines to identify homogenous groups of objects based on similarities. In machine learning with the increased generation of data, classification continues to be a key subject. While several literary works are interested in classifying the single label, the enormous dimensions of the data require a new approach. Multi-label clustering has therefore gained considerable attention in the testing community in recent years. This method involves a data instance with different labels and it is useful for many fields, e.g. image analysis, text classification and Big Data privacy security. In this case the classification of the single label is expanded. The high dimensionality of the distributed system needs an efficient and effective data management. Multi Label Classifier divides one or more labels in a set of labels of a particular instance. Multi-label classification is one of the leading data collection methods, where a set of labels is annotated in the data collection for each single instance. In one instance, the nature of multiple labels requires more computer power than classified one-label tasks. A multi-label grouping is often simplified by the method of splitting into one label classification, which avoids the distinction between labels. A Multi-Label Big Data Clustering with Privacy Protection Probability Linked Weight Optimization (MLBDC-PP-LWO) model is provided in this paper. In this proposed work, after the identification of sensitive data from data clusters, sensitive information is protected or generalized. The models proposed are compared to existing models and the findings show that the proposed model privacy preservation levels are more than the traditional methods.

Keywords: Unsupervised Data, Classification, Probabilistic Multi Labelled Data, Clustering Model, Linked Weight Optimization, Privacy Preservation, Big Data.

1. Introduction

As Internet technology advances rapidly, vast quantities of textual data are permanently generated due to large textual data by different fields. In reality, conventional machine learning techniques cannot be categorized, hence extracted from these data and considered as knowledge. Another problem is that most of the data cannot belong to a single group or label [1]. The data will certainly be related to one or more categories and this data is called multi-label data. Multi-label data classification is regarded as the classification of multi-label or multi-label learning. While multi-label classification is a very serious challenging problem in classification, it has received growing attention in the research field because of its high value in application; a text document can for example fall into one of the subsequent categories in the economic and computer categories.

There are three major types of categorization, binary, multi-class and multi-label classification, according to the Machine learning literature [2]. This categorization is focused primarily on the number of labels assigned in the data set to each instance [3]. If instances of a collection that have two positive / negative effect are linked to the single label, they can be recognized as binary classification, while any instance with a single label from a small set of outcomes annotates a multi-label classification problem in a dataset [4].

Huge amounts of data are produced habitually from different fields, such as medical, financial, library, telephone, shopping documents and individuals [5]. It is proven helpful for data mining applications to broadcast these data. Such data, on the one hand, is better for businesses to decide on verification. Data security aspects, on the other hand, may prohibit data owners from providing data analysis information [6]. The owner of the data can present a solution that provides the double aim of privacy preservation and also the accuracy of data mining tasks, such as clustering and classification to distribute data while persevering privacy [7]. The optimization technique is applied for achieving better results that is depicted in Figure 1.

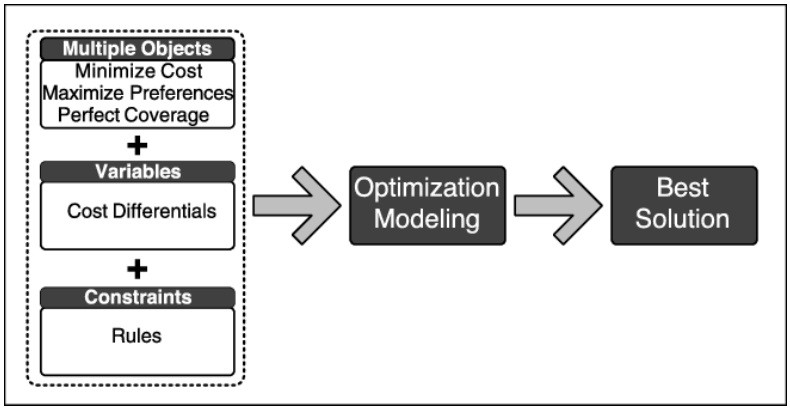

Fig. 1. Optimization Model

The big data refers to the vast volume of structured and unstructured data with a regular rise of 2.5 ExaBytes. This rapid growth in data volume was triggered by web services, mobile data, data on healthcare, GPS signals, Youtube and websites that host digital cameras and files and social media such as facebook and twitter. Big data can generally be categorized as volume, velocity and diversity. Volume is greater than tera bytes and peta bytes. Second, the speed of data in and out, data distribution and data appropriation is velocity [8]. Finally, diversity means an abundance of new forms of data from social locations, mobile computing and machinery. In companies and society, Big Data is used to develop big data analytics. However, the mixture of personal and international data makes sensitive attributes more vulnerable [9]. Prevention of confidential disclosure information is known as the security of big data [10]. Data violation is caused by large quantities of data, management, storage, handling and data analysis.

The primary issue in data security is the responsive nature of raw data. The data miner should not be able to access all sensitive information in its source form when collecting numerical information on the data. It does not always help to delete from the data set with simple strategies such as eliminating the exclusive personal identifier such as the name or social security identity, including personal information [11]. Because attacks on reidentification have increased, the link from various public data locations are not sufficiently secure to replicate the source subjects [12].

The division of data into different groups or clusters is called clustering, based on the similarities of the data. It is an unsupervised learning technique used to study interrelationships between a pattern set by grouping them into homogeneous clusters [13]. As similarity measure for clustering objects, relative distance or relative density between objects is considered [14]. Clustering takes place based on the theory to increase the resemblance of the intracluster and minimize the similarity of the intercluster [15]. In different areas, clustering is used. For instance, special catalogues are produced in a chain of department stores for different demographic groups based on attributes such as consumer sales, location and physical assets [16]. Clustering is carried out on the specified attribute values to decide the target mailing of the different catalogues and to help with the development of more detailed catalogues.

In order to balance the needs of scientific researchers and the privacy of individuals in their databases, the privacy module focuses on preservation [17]. Detection of privacy awareness is to detect how much a privacy assurance the device can protect data from the user, in order to compensate the communication between privacy and data usefulness [18]. Many of the datasets do not allow data subjects to be questioned for privacy purposes. In addition, it is not easy for data analysts to determine the degree of privacy assurance that they want to protect data security for their analysis results. This is because there is a need for considerable mathematical knowledge in the correct allocation of the minimal data protection strategies across various computations.

In this paper a Multi-Label Big Data Clustering with Privacy Protection Probability Linked Weight Optimization model is provided for performing clustering of big data and providing privacy for the data by performing multi labeling. The proposed model effectively generates the multi label clusters and the privacy rate of the proposed model is high.

2. Literature Survey

Data mining models for preservation of privacy can be split between data disturbance and safe multiparty calculation in two principal categories. In order to protect the confidential data of users and preserve critical

information, data disruption techniques are used to convert the original data. For data mining purposes, the transformed dataset is released. Two measurements are used to determine disruption techniques: privacy and information loss. In order to establish a disruption process, the protection of privacy is maximized and the information loss is minimized. Stable multi-party analytical methods are commonly used in distributed databases to preserve privacy. Data mining model measurement and extraction of useful information was performed in a distributed approach by sharing the minimum necessary information between the users participating without transferring the raw information.

Clustering is the mechanism used to divide the input space into a set of pre-specified classes. Similar objects are the essential property of the clustering. In several applications clustering was used, including user comfort ability analysis, bioinformatics and forensics, etc. A new field of research known as the clustering protection has arisen to resolve privacy concerns when data is shared. Clustering privacy guarantees the privacy and accurate clustering outcomes for individuals.

J. Yuan et al. [1] proposed a model to anonymize data set that contain confidential information, until they were unlimited for the mining industry, to consider the data mining applications. The databases are anonymized such that k-anonymity is preserved. Two general manipulation strategies are used, generalization and deletion, to achieve k-anonymity of a dataset. But generalization is also an important limitation, as it establishes a taxonomy of domain hierarchy for every value identifying in the data set to exercise k-anonymity. The expected well-organized, multidimensional deletion methods were carried out, i.e. values were only deleted on convinced records based on additional attribute values that do not require a physically constructed hierarchy of domains.

G. Wu et al [3] proposed a privacy classification method of data distortion in order to achieve privacy protection in data release. They addressed data mining algorithm assessment criteria like efficiency algorithm, data utilities, privacy security degrees, and data mining challenges. They proposed additional directions for the advancement of data protection mining. The hierarchy of classifications is indicated for analysis of data protection, classifying this method as the distribution of data, data change algorithm, knowledge or rule hiding and the preservation of privacy. They also discussed about different methods already existing in each methodology for classification. For various data mining techniques, they assessed algorithms linked to heuristic-based techniques, encrypted technologies and reconstruction-based techniques.

J. Ren et al. [5] considered the unimpeded access to data and to the information visualization process. The security of data was a primary priority for many legitimate data surveys worldwide. Some systems, like commercial data is richer in diagnostic values but are not communal at the same time without making any adjustments during the initial stages when reliable information is not used and the data is also unclean. Data mining models with different data conservation methods and successive data mining methods on such data is performed.

T. Shang et al [6] presented a random projection which, has high competence and usefulness, that has caused considerable concern among the privacy groups that protects data mining. With various hypotheses, the author opts for a numerous typical setting for the mining of data where the characteristics of the original data are communally autonomous and additional, a technique called maximum posteriori was given when the data features were linked and required, the author suggested the renovation techniques based on the Underdetermined Independent Component Analysis (UICA).

Q. Zhang et al. [8] introduced a data disruption framework for critical retention in k-clusters. In this case, geometrical transformations on the clusters disturbed the data objects into clusters using the k-means clustering. The entity of each cluster and the position of objects in the cluster remain the same throughout the k-means clustering. This geometric modification is accomplished during the rotation of the cluster, i.e. every cluster alternates around its center. The clusters are first relocated from the center of the entire dataset, so there is no relation between two clusters behind the resulting cluster rotation.

3. Proposed Method

Classification helps to classify the record in order to derive the data from particular aspects. Each record consists of a tuple of the collection of attributes and the mark of class. And also the tuples for characterization are represented by the predictor, independent and dependent variable. The improved data security framework in multi-partitioned information systems is structured efficiently to strengthen the privacy protection process for disrupted data in multi-partitioned data sets. The basic architecture of the proposed model is depicted in Figure 2.

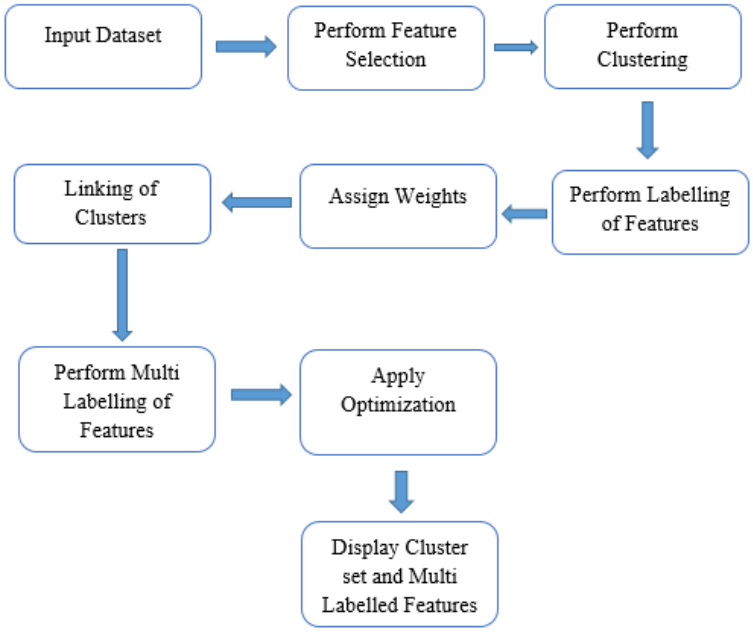

Fig. 2. Proposed Framework

The proposed Multi-Label Big Data Clustering with Privacy Protection Probability Linked Weight Optimization model initially performs clustering of data and then select relevant features from the clusters and assign weights to each and every feature extracted and perform labelling to those features in the initial level. The weights assigned features are then linked with other based on the weights and these features are labelled again and used for improving the privacy preservation levels by using optimization technique. The proposed model process is explained in the algorithm.

Algorithm MLBDC-PP-LWO

{

Input: Dataset(DS), Weight (W), Threshold (T), Labelling Token (LT)

Output: Cluster Set (CS), Multi Labelled Features (MLF)

For every instance in the dataset, perform clustering by comparing the values that are similar to form a cluster group as

For each I in DS(i)

$$F=\sum_{i=1} \frac{DS(i)+T+\lambda}{|LT|}$$

Where λ is the comparison difference of instances in the dataset.

The instances are compared with all the records in the set based on parameters and then grouped as a cluster using the equation

$$Cg(i) = F - \sum_{i \in DS(i)} \frac{|DS(i+1)^n|}{F(n)} * DS(n)+T$$

Where DS(n) is the last record and T is the Threshold limit.

After the clustering is performed, weights are calculated for the extracted features as

$$W(DS(i)) = \sum_{i=1}^{n} \frac{Cg(i) + \sum_{j=1}^{n} Cg(i+1) * (i-j)^2}{(i-j)^2 + T}$$

Where I and j are the neighboring features of a cluster. After Weights are calculated for each feature in a cluster, labelling is performed by assigning a labelling token LT as

$$L(Cg(i)) = \sum_{i \in DS(n), j=n-i} DS(j) W_j^l \left(F(i) + F(j)_i\right) + LT\left(i_j, j_i\right)$$

After labelling is performed to the features in the cluster group, all the cluster features are linked to for a Cluster Set CS. The linking of clusters is performed using the equation

$$\theta s_t(Cg(i)) = \sum_{i=1}^{N_i} L(W(i) \in DS(i) * L(Cg(i)) + \eta_{i,j}$$

Where $\eta_{i,j}$ represents the neighbor pixel similarity level. The similarity levels is calculated as

$$similarity(i,j) = \sum_i \sum_j |F - W| Max(W(i,j)| + T$$

$$L(DS(i), DS(n) = \sum_{i=1}^{N} \sum_{j=1}^{N_j} L(Cg(i) + \theta T_t(F - W) + \frac{|DS(i+1)^n|}{F(n) + W(F(i))}$$

Here θ represents the direction of linking of cluster groups. After all the cluster groups are linked, then labelling is again performed to achieve multi labelling clustering. The multi labelling is performed as

$$Cg(DS(i)) = \sum_i \sum_j \frac{(LT(i) - W_u)(DS(j) - W_v)}{\sigma_i \sigma_j}$$

$$Cg(DS(N)) = \sum_i \sum_j \frac{(L(DS(N-i)(i,j) - Cg_i Cg_j)}{\sigma_i \sigma_j + F(i)}$$

Final Labelling is done using the equation

$$MLF(W(DS(i)) = \sum_{i=i}^{N} \sum_{j=i-1}^{DS(N-J)} Cg(DS(i)) + \mathrm{Cg(DS(N)} - \mathrm{W(j)} * \mathrm{T}_i\,(LT - l)$$

After performing multi labelling features, optimization technique is applied with the equations

$$O\left(\frac{F_N}{DS(i)}, \theta\right) = \frac{\exp\left(F(i)_j{}^T W(i) + j_n\right)}{\sum_j \mathrm{Cg(DS(i)} + \exp\left(W(j)_i{}^T W(j) + i_n\right)}$$

The cost reduction is performed as

$$\mathrm{O(C(i))} = \frac{W(^x/_{\emptyset, F(i)_k}) + L(DS(i)) + T}{\sum_{i,j} \mathrm{MLF(DS(i)+DS(j)+xp(}^x/_{\theta T_t(F-W)}}$$

The proposed model considers only relevant features for providing the privacy preservation for improving the privacy levels of the model. The similarity of the features un the optimization technique is performed as

$$Sim(DS(i)) \varepsilon\, DS(i) = arg_{i,j}^{max} \sum_{i=1}^{N} w_{i,j} F(i) + \exp\left(O(C(i))_{i,j}{}^T + W(i)\right)$$

Finally, the Cluster set after performing multi labelling is performed as

$$ClusterSet(DS(i)) = \sum_i \sum_j ((MLF(i) + Cg(i)) + (MLF(j) + (Cg(j) - F(j)$$

Finally the cluster set is displayed with Multi labelling features and the cluster set and multi labelled features are displayed that provides high privacy levels to the users using the proposed model.

4. Results

The proposed model is implemented in python using ANACONDA SPYDER platform. The data sets are effectively partitioned by combinatorial feature either horizontally or vertically. The datasets are considered from the link https://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free. Multi Label Classifier divides one or more labels in a set of labels of a particular instance. Multi-label classification is one of the leading data collection methods, where a set of labels is annotated in the data collection for each single instance. In one instance, the nature of multiple labels requires more computer power than classified one-label tasks. A multi-label grouping is often simplified by the method of splitting into one label classification, which avoids the distinction between labels. In the proposed work, a Multi-Label Big Data Clustering with Privacy Protection Probability Linked Weight Optimization (MLBDC-PP-LWO) model is provided. Finally, the privacy and protection of disrupted data is protected through a method of selection in multi-labelled cluster sets. The

proposed model is compared with the traditional Trust Based K-Anonymity Clustering Method (TbKACM) in terms of privacy preservation levels, clustering time, data loss, Data Labelling Time, Accuracy.

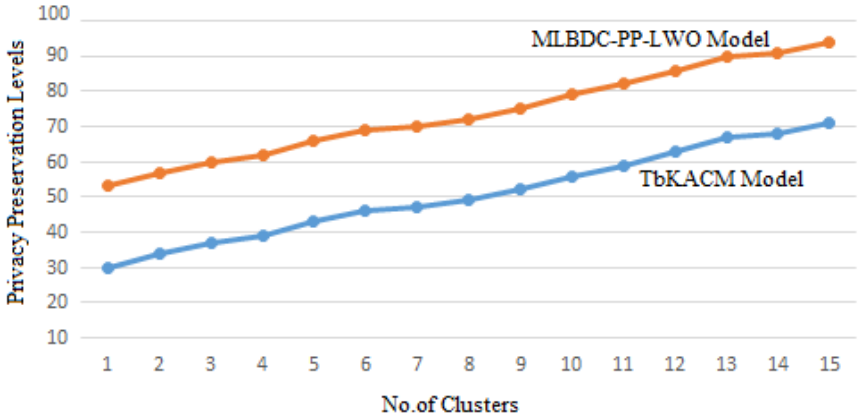

Fig. 3. Privacy Preservation Levels

The proposed model is compared with the existing Trust Based K-Anonymity Clustering Method (TbKACM) and the results in the Figure 3 shows that the privacy preservation levels of the proposed model is high than the traditional method.

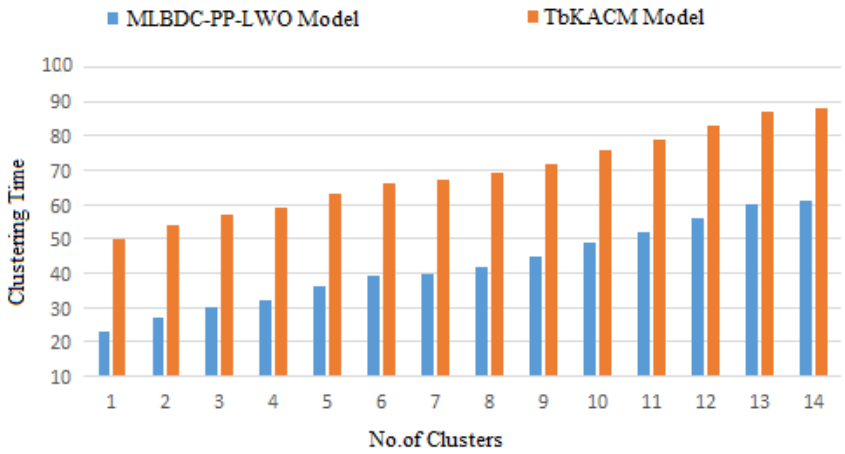

Fig. 4. Clustering Time

Figure 4 displays the clustering time levels of the proposed and existing methods. The proposed model is compared with the existing Trust Based K-Anonymity Clustering Method (TbKACM) and the results in the Figure 4 shows that the proposed model takes less time for clustering than the traditional method.

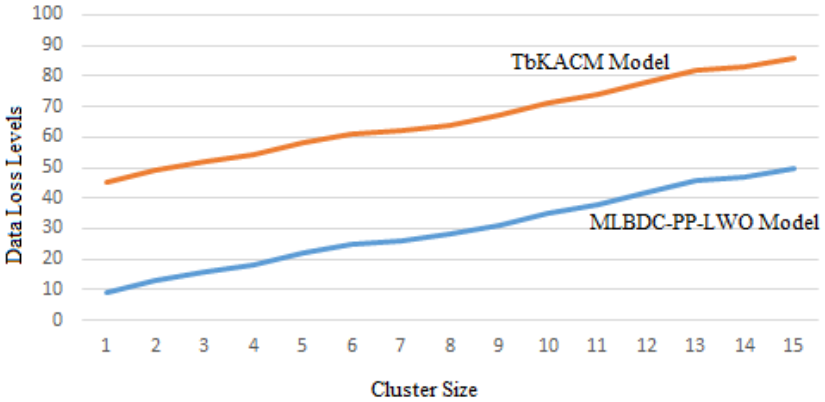

Fig. 5. Data Loss Levels

The data loss in the proposed model is very less when compared to the existing method. The Figure 5 illustrates the data loss levels in the proposed and traditional methods. The proposed model performs feature extraction using optimization technique. The user's data is kept secret which indicates the data loss is less in the proposed model.

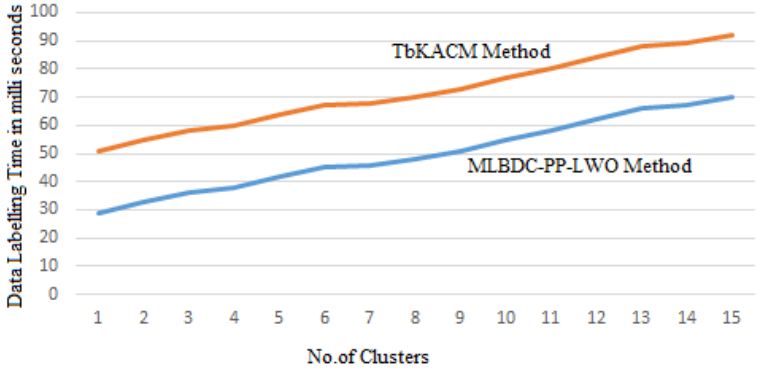

Fig. 6. Data Labelling Time

The Proposed model performs multi labelling of data features. The multi labelling time levels of data features in the proposed method is very less when compared to the existing method. The Figure 6 depicts the Data Labelling Time levels of the proposed and existing methods.

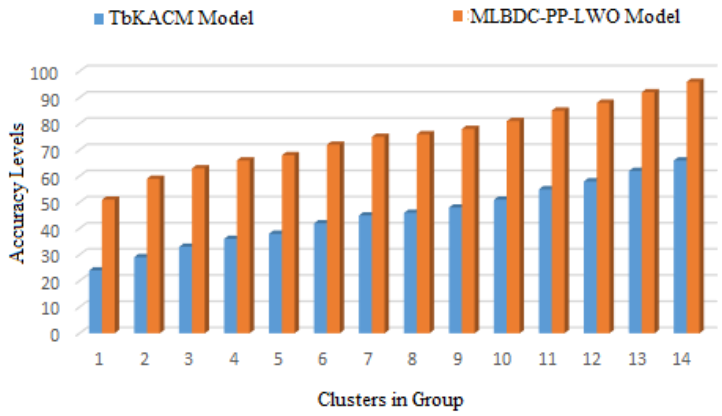

Fig. 7. Accuracy Levels

The Figure 7 depicts the accuracy levels of the proposed and traditional methods. The proposed model performs multi labelling of features and then optimization technique is used for reducing the cost and data loss. The proposed model improves the accuracy levels by using the multi labelling linked weight optimization technique.

5. Conclusion

The architecture of data security is a key issue and the partitioned data sets are efficiently exchanged using a linked weight optimization model. The partitioned data sets are clustered using linked weight optimization model to boost the privacy method effectively by performing multi labelling method. The proposed clustering scheme removed the complexity by changing the efficient cluster selection criterion to a given threshold over horizontally and vertically partitioned data set. Furthermore, the security of privacy is enhanced by choosing levels of privacy. In multiple data sets, the application of privacy preservation to unsupervised data effectively deliberates the privacy security process by subsequently analyzing the unsupervised data. Finally, the improved privacy conservation using linked weight optimization guarantees confidentiality and security. The partitioned data set is labelled with a set of features and classes that facilitate the collection and protection process. The benefit ratio value gathers feature for the optimum set of features to enhance data validity. Finally, in enhanced privacy with disturbed data, the privacy and protection of the disturbed data are strengthened using selection of functions. Improved data security with disrupted data can be further strengthened using feature selection by integrating knowledge of context as a potential research initiative. Context information can be used to guide the

process of finding and allow revealed patterns to be generated in concise terms at various levels to improve privacy preservation levels.

References

1. J. Yuan and Y. Tian, "Practical privacy-preserving map reduce based k-means clustering over large-scale dataset," *IEEE Transactions on Cloud Computing,* vol. 7, no. 2, pp. 568–579, 2019.
2. H. Yan, J. Li, and Y. Zhang, "Remote data checking with a designated verifier in cloud storage," *IEEE Systems Journal,* pp. 1–10, 2019.
3. G. Wu, Y. Mu, W. Susilo, F. Guo, and F. Zhang, "Threshold privacy-preserving cloud auditing with multiple uploaders," *International Journal of Information Security,* vol. 18, no. 3, pp. 321–331, 2019.
4. J. Li, H. Yan, and Y. Zhang, "Certificateless public integrity checking of group shared data on cloud storage," *IEEE Transactions on Services Computing,* pp. 1–10, 2018.
5. J. Ren, J. Xiong, Z. Yao, R. Ma, and M. Lin, "Dplk-means: a novel differential privacy k-means mechanism," *In Proceedings of the 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC),* pp. 133–139, Shenzhen, China, June 2017.
6. T. Shang, Z. Zhao, Z. Guan, and J. Liu, "A DP canopy k-means algorithm for privacy preservation of hadoop platform," *In Proceedings of the CSS 2017, Lecture Notes in Computer Science,* vol. 10581, pp. 189–198, Springer, Xi'an, China, October 2017.
7. H. Rong, H. Wang, J. Liu, J. Hao, and M. Xian, "Outsourced k-means clustering over encrypted data under multiple keys in spark framework," *In Proceedings of the Secure Comm 2017, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering,* vol. 238, pp. 67–87, Springer, Niagara Falls, ON, Canada, October 2017.
8. Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "Pphopcm: privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing," *IEEE Transactions on Big Data,* 2017.
9. G. Wu, Y. Mu, W. Susilo, and F. Guo, "Privacy-preserving cloud auditing with multiple uploaders," *Information Security Practice and Experience,* vol. 10060, pp. 224–237, 2016.
10. M. Du, K. Wang, Y. Chen, X. Wang, and Y. Sun. "Big Data Privacy Preserving in Multi-Access Edge Computing for Heterogeneous Internet of Things". *In: IEEE Communications Magazine* (2018), pp. 62–67.
11. M. Abadi, A. Chu, I. Goodfellow, H. Brendan McMahan, I. Mironov, K. Talwar, and L. Zhang. "Deep Learning with Differential Privacy". *In: ArXiv e-prints* (July 2016).
12. Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep Learning with Differential Privacy". *In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. CCS '16.* Vienna, Austria: ACM, 2016, pp. 308–318.
13. Australian National Data Service (ANDS). Application process to research on sensitive data with Ethics and Consent. https://www.adrn. ac. uk / get - data / application - process/ANDS's application process and https://utas.libguides.com/researchdatamanagement/ ethics_sensitivedataANDS's ethics and consent. 2017. (Visited on 06/30/2017).
14. Mohit Bansal, Kevin Gimpel, and Karen Livescu. "Tailoring Continuous Word Representations for Dependency Parsing". *In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Baltimore, Maryland: Association for Computational Linguistics,* 2014, 809–815.
15. E. Barendt. Privacy. *The International Library of Essays in Law and Legal Theory (Second Series).* Taylor & Francis, 2017. isbn: 9781351908801.
16. Kayhan N. Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Samuel Gershman. "Nonparametric Spherical Topic Modeling with Word Embeddings". *In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (2016), pp. 537–542.
17. Bhume Bhumiratana and Matt Bishop. "Privacy Aware Data Sharing: Balancing the Usability and Privacy of Datasets". *In: Proceedings of the 2Nd International Conference on PErvasive Technologies Related to Assistive Environments. PETRA '09. Corfu, Greece: ACM,* 2009, 73:1– 73:8.
18. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information". *In: Transactions of the Association for Computational Linguistics 5* (2017), pp. 135–146.