International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Comprehensive Research on Privacy Preserving Emphasizing on Distributed Clustering

Prajna M.S.[1], Sumana M.[2]

[1, 2]Department of ISE, M.S. Ramaiah Institute of Technology, Bangalore, India

Abstract: ***Often, the information is sensitive or private in nature and these sensitive data when mined violates the privacy of the individuals. Privacy preserving data mining (PPDM) mines the data but intends to preserve the privacy of susceptible data without ever actually seeing it. This paper recaps the important techniques in PPDM like anonymization, perturbation and cryptography. Nowadays, data mining is extensively used when the data is distributed among multiple parties. This paper highlights the research carried out in privacy preserving distributed clustering. Clustering is an effective method to discover data distribution and patterns in datasets. Significant research in privacy preserving distributed clustering is shaped on k-means clustering algorithm with secure multiparty computation (SMC). This work focuses on the previous development, existing challenges, and upcoming trends in privacy preserving k-means clustering with horizontally and vertically distributed data.***

Keywords: privacy preserving data mining, distributed data, k-means clustering, secure multiparty computation

1. Introduction

In recent years, there is huge advancement and development in network technologies, internet, computing applications as well as data mining programs. The governments, corporations, organizations and individuals collect a large volume of digital information. This has brought great opportunities in the research area of data mining. Data mining technology discovers patterns, associations, performs classification and prediction from data [1]. Several applications like social networking, medical domain and banking utilize the data mining technology to analyze the data from various viewpoints and summarize it into worthwhile information. Many organizations accumulate and hold a bulk capacity of data which is further processed by data mining tools to raise the revenue or lessen the costs or both [2].

In numerous circumstances data is private or valuable, where data possessors hesitate to disclose their sensitive information. Alternatively, the individual's private data if published or disclosed may raise legal, ethical or privacy concerns. With the increase in technology, it is easy to retrieve and mine precise data which is linked to other datasets. The traditional data mining technology and algorithms directly function on the original dataset which may lead to the leakage of private data. As a result, the research area of data mining considers privacy as a significant concern [3].

Privacy preserving data mining (PPDM) is an important area to study in data mining which deals with various privacy issues. PPDM is a duo of data mining and information security ensuring data privacy [3], [4]. Thus, PPDM allows data mining algorithm to process data without ever actually seeing it. So, private data remain private even after the mining process. The PPDM technique aims at the following.

- It should prevent the extraction of confidential information.
- Data mining techniques should not be compromised.
- The computational complexity must not be high.
- It should ensure non-disclosure of private data.

The organization of the paper is as follows: Section 2 highlights the concepts in privacy preserving data mining. Section 3 explains distributed clustering in privacy preserving, emphasizing on the k-means algorithm. The concept of secure multiparty computation is explained in Section 4. The various research approaches in privacy preserving distributed clustering along with data distribution is captured in section 5. At last, section 6 concludes the article with future directions.

2. PPDM: A Study

Privacy preserving data mining is an interesting research zone which explores how the privacy of data can be upheld either before or after applying data mining methods to the data. In the field of data communication and data mining or knowledge discovery, privacy preservation of sensitive or private information is an important area which needs special attention. The pioneering work in the arena of privacy preserving was instigated by Agarwal et al. [3] and Lindell and Pinkas [5]. The objective of their work was to mine facts from users' personal data without revealing the data items.
The next section describes the privacy preserving data mining framework and its classifications.

2.1 PPDM Framework

The three level frameworks for privacy preserving data mining are depicted in Figure 1. The first level is a transaction phase. In this level, raw data is gathered from single or multiple sources or systems. These collected data is then preprocessed or transformed to fit into the desired format and is kept in a data warehouse. In the second phase, data mining algorithms are applied on data to infer the knowledge. It is ensured that data mining algorithms are adapted so that it protects privacy without giving up its main intention. The third phase is the output phase where results are generated [6].

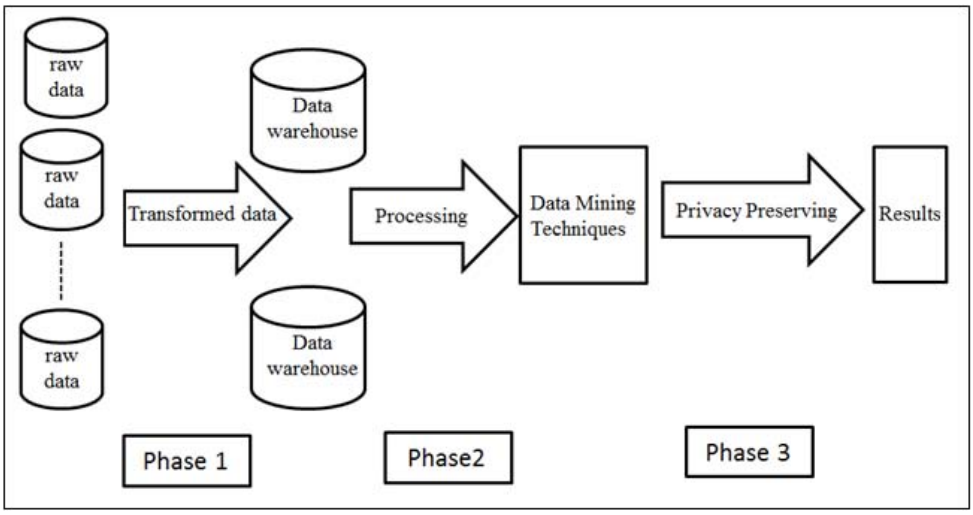

Figure 1: Flow of PPDM

2.2 PPDM classification hierarchy

Most of the time, the application using data mining guess that the data is from a single central source data mining or data warehouse. If single repositories data is violated then entire data may be revealed which is a threat to privacy. One of the ways to escape from the above scenario is to avoid a centralized warehouse and construct a distributed system which reduces the data exchange [7]. Figure 2 demonstrates the centralized and distributed scenario of classifying PPDM and the techniques used. The main focus of this paper is distributed scenario which is explained in the next section.

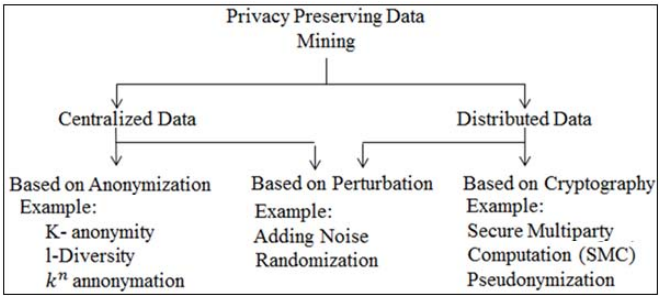

Figure 2: PPDM classification on data distribution

Anonymization techniques as the name suggests, covers the individuals sensitive information among group records either by removing or encrypting [8]. Data removal technique, data generalization technique, data suppression technique, swapping technique, permutation technique can be exploited to achieve anonymization. Generalization technique replaces the value with semantically reliable value whereas; suppression technique blocks the value. The conventional model in anonymization is k-anonymity [9] which leads to several further types of researches. The work [10] describes the improved versions of this technique such as km-anonymization, l-diversity, t-closeness, etc.

Data perturbation method distorts the original data so as it does not reflect true values which ensure the privacy. Data swapping is one of the perturbation techniques. Here the data values are exchanged or swapped between the records so as to preserve privacy. Alternatively, another technique is randomization which adds noise to data to hide real data [11]. Both centralized and distributed scenarios cope up with perturbation technique. The paper [12], [13] describes the approaches used in distributed settings. Perturbation technique impacts the result of data mining which has to be solved. The work in [12] introduces to study carried out on geometric data. The paper [13] is constructed on Principal Component Analysis, it clusters the distributed datasets.

In the distributed scenario, cryptographic techniques are widely used. The paper [3] is on distributed ID3 which is an efficient privacy preserving protocol. The work in [5] describes the cryptographic techniques for PPDM which is based on Oblivious Transfer. Pseudonymization and SMC are the two main techniques in cryptography based systems. Pseudonymization is a combination of anonymization and data encryption. This technique uses pseudonyms by which linking two records become un-linkable. It is a reversible technique which requires encrypting only metadata so computational overhead can be decreased [14].

Consider the scenario of revealing private data to the referring parties. This is not a breach of privacy whereas; revealing information to other parties is a breach. Secure multiparty computation (SMC) is such a technique which can ensure privacy in distributed environments. SMC includes homomorphic encryption, secret sharing and circuit evaluation [15]. Section 4 explains SMC in more detail with respect to distributed clustering. Table 1 summarizes the different techniques used along with its strengths and weakness.

Table 1: Summary of PPDM techniques

	Techniques	*Settings*	*Strength*	*Weakness*
Anonymization	Suppression, Permutation, Generalization	Centralized	Identity non-disclosure	Less accurate, Loss of data
Perturbation	Swapping, Adding noise	Centralized, Distributed	Easy, Efficient	Loss of data
Cryptography	Cryptographic based	Centralized, Distributed	Preserves information, Well defined approach	Complex computation

3. Privacy Preserving Distributed Clustering

Clustering is an empirical unsupervised learning process in data mining. Clustering process learns data distribution and patterns in underlying datasets. Clustering can be defined as the method of organizing entities into meaningful clusters whose members are similar in some way [16]. The important applications of clustering are in the fields of social networking, medical domain and banking, marketing, climate. In the scenario of distributed datasets, the clustering method is performed on the aggregated datasets from multiple parties.

There exist plenty of techniques and algorithms for clustering. Each algorithm has its own strengths and weaknesses, so it is used rendering to the essential application and data type. In the field of data mining, there are many clustering methods like hierarchical methods, density based, partitioning based, model based methods etc. There exists a partitional, fuzzy, hierarchical or nested algorithm in clustering [17], [18]. The prominent algorithm in this class is k-means clustering [4], [19]. This paper illuminates the k-means clustering algorithm and the research carried out regarding distributed data.

3.1 K-means clustering Algorithm

K-means clustering is a method to group records or items into k clusters such a way that items fits into the cluster through the closest center [20], [21]. The k-means clustering begins with initializing the desired number of clusters that needs to be designed for the datasets. Later k imitation objects are formed as the initial cluster centers which are allotted randomly.

Further, the algorithm continues by interchanging between assignment and update steps. In assignment step, the distance from k cluster centers to every record is calculated and each object is allocated to the nearest mean. To compute the distance matrix Euclidean, Manhattan, or Minkowski can be used [22]. In the update stage, the cluster centers are updated in accordance with the records in each cluster. The steps are recurred until there is no change in cluster centers or until the desired iteration is reached.

Let's assume there are n data points say $x_1, x_2, \ldots, x_n$ of real numbers in m dimensional vector. Let 'k' be the number of desired clusters. Initialize, $c_1, c_2, \ldots, c_k$ as the new cluster centers. Then the k-means algorithm proceeds as follows.

1. Randomly select 'k' cluster centers.
2. Calculate the distance between each cluster centers and every data point.
3. Repeat
   a) Assignment Step: Assign each data entity to the nearest cluster center
   b) Update Step: Update the centers for each cluster
4. Until there is no variation amongst old and new cluster centers.

The distance between cluster center c_i and data entity x_i can be calculated using the equation (1).

$$\sum_{i=1}^{n} \left(c_i - x_i \right)^2 \quad (1)$$

The cluster centers c_i having $\{x_1, x_2, \ldots, x_n\}$ can be updated using the equation (2).

$$v_i = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad (2)$$

3.2 Distributed k-means clustering

In distributed clustering [2, 23], the dataset is partitioned among n parties. Each party desires to mutually cluster their datasets without leaking any sensitive information. Assume that party A has datasets$\{x_1, \ldots, x_a\}$, and party B has datasets $\{x_{a+1}, \ldots, x_b\}$ and party C has $\{x_{b+1}, \ldots, x_c\}$ datasets and so on. It is noble to cluster the combination of datasets of the multiple parties than clustering the individual data set and then aggregating the result. The update step explained in the k-means algorithm needs to be modified for distributed dataset. In the scenario of distributed clustering with the presence of a trusted third party (TTP), multiple parties locally implement the computations. The new cluster centers are calculated by interacting with the TTP.

4. Secure Multiparty Computation (SMC)

In the scenario of secure multiparty computation (SMC), data is distributed among multiple distinct yet connected parties. All parties cooperate to securely compute the final result without revealing their individual data to any party. SMC can be used as a cryptographic alternate to a trusted third party (TTP) as TTP is expensive and is hard to find one. The idea of secure two-party computation was initiated by Yao [24].It was later extended in the work by [25] to fit into multiple parties. There are two types of adversaries in SMC. Passive adversary is also known as semi-honest model. Though, it satisfies the directions with its correct input, it has the liberty to use what it sees during the execution. The other adversary is malicious model [26].

Privacy and correctness are the vital necessities of any secure computation. Privacy holds when parties know only their output nothing outside what is undeniably required. When each party obtains its correct output correctness holds good. For the developed protocol security must be ensured. In terms of cryptography, the protocol is secure if the protocol exhibits probabilistic nature. The secure multiparty computations uses cryptographic or randomization method. In randomization, noise is added to the data. This prevents identifying the real data [9]. The next section briefs out important techniques in secure multiparty computations.

4.1 Homomorphic Schemes

The homomorphic property allows operating on cipher values and obtains cipher value as results. In privacy preserving clustering technique, homomorphic public key cryptosystem is used to securely compute the distances matrix and cluster centers. The two main homomorphic techniques used in privacy preserving distributed clustering

are homomorphic encryption and secret sharing [27]. Homomorphic encryption is public key encryption which is semantically secure. It satisfies the following property.

$$E(m_1)\cdot E(m_2) = E(m_1 + m_2) \quad (3)$$

$$E(m_1)^{m_2} = E(m_1 \cdot m_2) \quad (4)$$

Where m1 and m2 are the texts and E is the function for public key encryption. One of the widely used homomorphic technique for distributed clustering is Paillier encryption [28].

4.2 Secret Sharing

Secret sharing is another very popular homomorphic scheme used in clustering distributed datasets. Initially, Shamir and Blakley [29] used this technique. Here the secret of one party is distributed amongst other parties in such a way that recovering the secret by one party all alone is not possible. There should be minimum threshold number of parties say t among n parties to gather the secret. The retrieval of secret fails if less than t parties try to find the secret. This technique is additively homomorphic [30].

4.3 Circuit Evaluation

The concept by Yao [23] is the fundamental work for evaluation circuit. Many encryption protocols in privacy preserving use this concept [27], [30]. Here a scrambled boolean circuit consisting of encryption values is used for function evaluation. Here the input is divided between the parties. The main advantage of this circuit is that the parties cannot learn anything apart from the result. As each bit requires encryption, this method is expensive. This approach is widely used in k- means clustering algorithm to securely find the distance.

5. Research on k-means distributed clustering: Review

K-means clustering algorithm is widely accepted technique used for clustering huge datasets in privacy preserving data mining [27]. This section brings out the various research works carried out on k-means clustering over different data distribution. Nowadays, databases are distributed among two or more parties. In Privacy-Preserving Distributed Data Mining (PPDDM) multiple participants jointly achieve a data mining task on the private data of all the participants ensuring that the participants have no knowledge of others data [31].This scenario fetches the aggregated result on the multiparty datasets.

Data partitioning or distribution is mainly categorized into horizontal and vertical distribution. There are other models exists as well, one of them is the combination of horizontal and vertical partitioning called as arbitrary partitioning [7]. In the distributed environment, multiple parties virtually have their combined data. Further, the imaginary database D consisting of z records is represented by D = {d_1, d_2, . . . , d_z} where di has n values for n attributes ($d_{i,1}$, $d_{i,2}$, … , $d_{i,n}$).

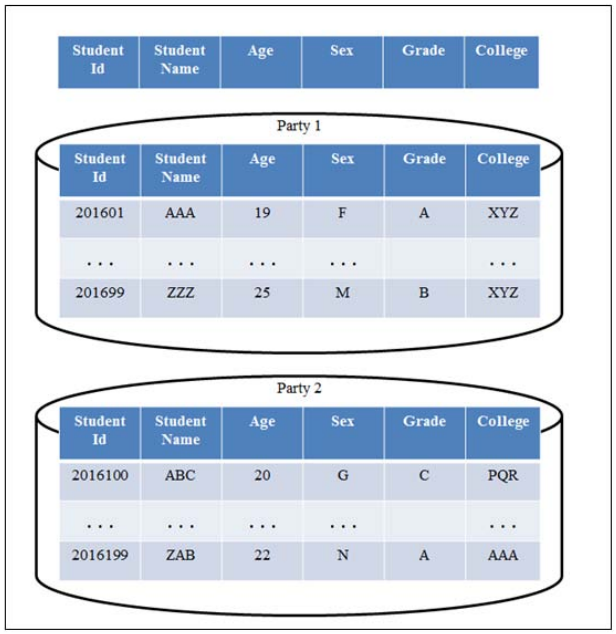

Figure 3: Horizontal data distribution

Horizontally partitioned data is also known as homogeneous distribution [32]. In this scenario, multiple parties have a similar type of information for different entities. Figure 3 illustrates horizontal partitioning among three parties. Consider the example of multiple parties holding student datasets. Here each party has student data belonging to different colleges. The imaginary database contains whole student database from all colleges.

In heterogeneous partitioning or vertical partitioning of data, multiple parties have different facts of the same set of objects [7]. Fig 4 depicts the scenario of vertically partitioned datasets between two parties, where Party1 holds patient information while Party 2 has bank details to the corresponding individual. As the dataset is distributed across multiple locations, the k-means algorithm requires modification depending on the data distribution. The next section depicts the work carried out in k- means clustering over data distribution.

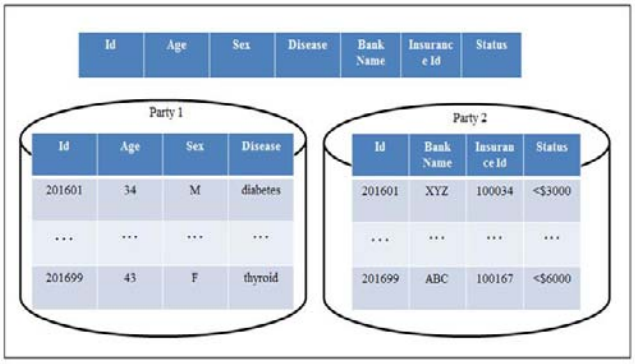

Figure 4: vertical data distribution

5.1 Research on k-means horizontally distributed clustering

In the scenario of horizontally distributed datasets, all the components of a record are present in every party. So, the

distance computation to cluster centroids will not breach privacy. But, the count of objects in each cluster is required to update the cluster centers which may lead to privacy risk. Another threat in horizontal distribution is randomly choosing the initial cluster centers. In this scenario, care must be taken to disclose the intermediate cluster center among multiple parties.

The research by Jha et al. [33] is one of the revolutionary works in privacy preserving data mining. In this work author proposes two protocols for clustering the horizontally distributed data based on k-means algorithm. This scenario demonstrates two parties participating in the clustering activity. The very first approach termed OPE is constructed on oblivious polynomial evaluation whereas the second approach termed DPE is assisted by homomorphic schemes. With respect to communication cost and computing, DPE appears to be efficient. The protocol is setup on semi-honest advisory model. This technique involves locally computing the distance between centroids and entities. Here, though the datasets of each party is kept private, the cluster centers are disclosed.

Saeed Samet et al. [22] have proposed a way out to privacy preserving clustering in the multiparty environment. This protocol does not reveal the count of entities in each cluster, but always discloses the intermediate cluster centers. Sankita et al. [23] used elliptic curve cryptography to preserve privacy over horizontally distributed datasets. This technique uses semi-honest advisory model. In this approach, multiple party uses ring topology to communicate among them.

Table 2: Summary of works in k-means horizontally distributed clustering

Author	*Number of parties*	*Techniques used*
Jha et al. [33]	2	Homomorphic encryption, Oblivious polynomial evaluation
Samet et al. [22]	n	Secure sum, Secure multiparty addition
Sankita et al. [23]	n	Elliptic curve cryptography, Homomorphic encryption, Ring topology

5.2 Research on k-means vertically distributed clustering

In the scenario of vertically distributed clustering, multiple parties have different facts of the same set of objects. So, there is no threat to privacy by the selection of initial cluster centers. In the case of computing the distance between entities to centroids, there might be a hint of data exposure. One more threat is while assigning each entity to the closest cluster. There might be a chance of exposing the total count of entities in each cluster while updating the centroids.

Vaidya et al. [34] first provided the solution to vertically distributed data. This protocol involves multiple parties and built on the semi-honest model. The concept of secure permutation and the homomorphic scheme is used to find the distance between objects in multiple parties. This protocol uses Yao's evaluation circuit [24]. Here, the total object count in each cluster as well as intermediate cluster values is revealed.

Doganay et al. [35] followed the approach of Vaidya with additive secret sharing instead of homomorphic technology. This approach used four non-colluding sites to lower the computation and communication cost. Samet bet al. [22] proposed another approach for clustering vertically distributed data using secure sum and secure multiparty addition. In the work of Zekeriya Erkin et al. [2] used a service provider, who performs clustering without a decryption key and also does not learn anything about the content of sensitive data. Here homomorphic encryption for distributed environment is used for protecting subtle intermediary values and the ultimate clustering tasks.

Table 2: Summary of works in k-means vertically distributed clustering

Author	*Number of parties*	*Techniques used*
Vaidya J et al. [34]	n>2	Secure permutation, Yao evaluation circuit Homomorphic encryption
Doganay et al.[35]	n>3	Additive secret sharing
Samet et al. [22]	n	Secure sum, Secure multiparty addition
Erkin et al.[2]	n	Encryption, Homomorphic schemes

6. Conclusion

In this contemporary world, preserving privacy of knowledge mining is very significant. This article presents a general outline on popular PPDM techniques and highlights their strength and weakness. The studies indicate the tradeoffs between privacy, accuracy, security, information loss and overhead in computation. This paper focuses on remarkable work in privacy preserving clustering with horizontal and vertical data distribution. In the scenario of k-means clustering, rigorous care must be taken to protect the information like intermediate cluster assignments, count of entities in each cluster, cluster centers. Finally, vital secure rules are desired to implement distributed clustering to preserve the privacy of multiple parties.

References

[1] Fayyad, Usama M. "Data mining and knowledge discovery: Making sense out of data." IEEE Intelligent Systems 5 (1996): 20-25.
[2] Erkin, Zekeriya, et al. "Privacy-preserving distributed clustering." EURASIP Journal on Information Security 2013.1 (2013): 1-15.
[3] Agrawal, Rakesh, and Ramakrishnan Srikant. "Privacy-preserving data mining." ACM Sigmod Record. Vol. 29. No. 2. ACM, 2000.
[4] Vaidya, Jaideep, Christopher W. Clifton, and Yu Michael Zhu. Privacy preserving data mining. Vol. 19. Springer Science & Business Media, 2006.

[5] Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining."Advances in Cryptology—CRYPTO 2000. Springer Berlin Heidelberg, 2000.
[6] Prakash, Mangal, and G. Singaravel. "A new model for privacy preserving sensitive data mining." Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on. IEEE, 2012.
[7] Vaidya, Jaideep, and Chris Clifton. "Privacy-preserving data mining: Why, how, and when." IEEE Security & Privacy 6 (2004): 19-27.
[8] Zhou, Bin, Jian Pei, and WoShun Luk. "A brief survey on anonymization techniques for privacy preserving publishing of social network data." ACM Sigkdd Explorations Newsletter 10.2 (2008): 12-22.
[9] Domingo-Ferrer, Josep, and Vicenç Torra. "A critique of k-anonymity and some of its enhancements." Availability, Reliability and Security, 2008. ARES 08. Third International Conference on. IEEE, 2008.
[10] Zhu, Yan, and Lin Peng. "Study on k-anonymity models of sharing medical information." Service Systems and Service Management, 2007 International Conference on. IEEE, 2007.
[11] Chen, Keke, and Ling Liu. "A survey of multiplicative perturbation for privacy-preserving data mining." Privacy-Preserving Data Mining. Springer US, 2008. 157-181.
[12] Chen, Keke, and Ling Liu. "Privacy-preserving multiparty collaborative mining with geometric data perturbation." Parallel and Distributed Systems, IEEE Transactions on 20.12 (2009): 1764-1776.
[13] Banu, R. Vidya, and N. Nagaveni. "Evaluation of a perturbation-based technique for privacy preservation in a multi-party clustering scenario."Information Sciences 232 (2013): 437-448.
[14] Claerhout, Brecht, and G. J. E. DeMoor. "Privacy protection for clinical and genomic data: The use of privacy-enhancing techniques in medicine."International Journal of Medical Informatics 74.2 (2005): 257-265.
[15] Franklin, Matthew, and Moti Yung. "The varieties of secure distributed computation." Sequences II. Springer New York, 1993. 392-417.
[16] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.
[17] Jha, Somesh, Luis Kruger, and Patrick McDaniel. "Privacy preserving clustering." Computer Security–ESORICS 2005. Springer Berlin Heidelberg, 2005. 397-417.
[18] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.
[19] Jagannathan, Geetha, and Rebecca N. Wright. "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.
[20] Bunn, Paul, and Rafail Ostrovsky. "Secure two-party k-means clustering."Proceedings of the 14th ACM conference on Computer and communications security. ACM, 2007.
[21] Oliveira, Stanley RM, and Osmar R. Zaïane. "Achieving privacy preservation when sharing data for clustering." Secure Data Management. Springer Berlin Heidelberg, 2004. 67-82.
[22] Samet, Saeed, Ali Miri, and Luis Orozco-Barbosa. "Privacy Preserving k-Means Clustering in Multi-Party Environment." SECRYPT. 2007.
[23] Patel, Sankita J., Dharmen Punjani, and Devesh C. Jinwala. "An Efficient Approach for Privacy Preserving Distributed Clustering in Semi-honest Model Using Elliptic Curve Cryptography." International Journal of Network Security17.3 (2015): 328-339.
[24] Yao, Andrew C. "Protocols for secure computations." Foundations of Computer Science, 1982. SFCS'08. 23rd Annual Symposium on. IEEE, 1982.
[25] Goldreich, Oded. Foundations of cryptography: volume 2, basic applications. Cambridge university press, 2009.
[26] Lindell, Yehuda, and Benny Pinkas. "Secure multiparty computation for privacy-preserving data mining." Journal of Privacy and Confidentiality 1.1 (2009): 5.
[27] Meskine, Fatima, and Safia Nait Bahloul. "Privacy preserving k-means clustering: a survey research." Int. Arab J. Inf. Technol. 9.2 (2012): 194-200.
[28] Paillier, Pascal. "Public-key cryptosystems based on composite degree residuosity classes." Advances in cryptology—EUROCRYPT'99. Springer Berlin Heidelberg, 1999.
[29] Dawson, Ed, and Diane Donovan. "The breadth of Shamir's secret-sharing scheme." Computers & Security 13.1 (1994): 69-78.
[30] Pedersen, Thomas Brochmann, Yücel Saygın, and Erkay Savaş. "Secret charing vs. encryption-based techniques for privacy preserving data mining." (2007).
[31] Zeng, Yong, et al. "Secure collaboration in global design and supply chain environment: Problem analysis and literature review." Computers in Industry63.6 (2012): 545-556.
[32] Liu, Jinfei, et al. "Privacy preserving distributed DBSCAN clustering."Proceedings of the 2012 Joint EDBT/ICDT Workshops. ACM, 2012.
[33] Jha, Somesh, Luis Kruger, and Patrick McDaniel. "Privacy preserving clustering." Computer Security–ESORICS 2005. Springer Berlin Heidelberg, 2005. 397-417.
[34] Vaidya, Jaideep, and Chris Clifton. "Privacy-preserving k-means clustering over vertically partitioned data." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.
[35] Doganay, Mahir Can, et al. "Distributed privacy preserving k-means clustering with additive secret sharing." Proceedings of the 2008 international workshop on Privacy and anonymity in information society. ACM, 2008

Author Profile

Prajna M.S. received her B.E (CSE) in 2009 from K.V.G. College of Engineering, Sullia, affiliated to VTU Belgaum. She is currently pursuing M.Tech in Software Engineering from MSRIT affiliated to VTU,

Belgaum. Her research interests include data mining, cloud computing and big data.

M. Sumana received her B.E (CSE) from M.I.T, Karnataka, in 2000, and her M.Tech. degree from VTU University in 2007, Karnataka and is currently pursuing the Ph.D. degree in privacy preserving data mining from the Manipal University, Karnataka, India. She is presently working as an assistant professor in the Department of I.S.E in M. S. Ramaiah Institute of Technology since 2007.