Privacy Preserving K-means Clustering: A Survey Research

Fatima Meskine[1] and Safia Nait Bahloul[2]
[1]Department of Computer Science, University of Es-Senia, Algeria
[2]Department of Computer Science, University Oran, Algeria

Abstract: *The clustering is an exploratory task of data mining. This raised a real problem of privacy when the data are from different sources. Most of researches on privacy preservation in clustering are developed for k-means clustering algorithm, by applying the secure multi-party computation framework. The distribution of data may be different (vertical, horizontal or arbitrary). Approaches allowing solving the problem on a vertical, horizontal and even arbitrary partitioned dataset were proposed. The major interest is to reveal the minimum of information during the execution of the algorithm, especially in k-means iterations, which poses a real challenge for secure multi party computation. This work consists to study and analyze all works of privacy preserving in the k-means algorithm, classify the various approaches according to the used data distribution while presenting the weaknesses and the strong points of each protocol regards to privacy. The interest is to arise the real needs of privacy during the execution of the different steps of k-mean algorithm, thus to discover the best of approaches in case of preserving privacy in k-means algorithm.*

Keywords: *k-means clustering algorithm, secure multi-party computation, distributed data, privacy preserving.*

Received December 31, 2010, accepted March 1, 2011

1. Introduction

The very active research area of privacy preserving data mining aims to extract useful information from data coming from multiple sources, while preserving these data against disclosure or loss. The first works of privacy preserving data mining were given by [2, 19] on the ID3 algorithm for decision trees classification [24]. Each work used a different model of privacy. In [2], privacy is preserved by randomizing the original dataset, and then provides it as input to the algorithm. In [19], a more rigorous model is used, which is a secure multi - party computation [11, 12]. In this approach, cryptographic primitives are added to the ID3 algorithm in order to protect distributed data from disclosure when executing it. Since, more researches based on secure multi –party computation for different algorithms of data mining were established [8, 16].

Lately, many studies were interested by views and surveys on privacy preserving data mining works [1, 30, 32], in attempt at generalization the results for different categories of data mining methods. We believe that the diversity of data mining methods and their specific use, limits the applied privacy preserving models to these methods. To better control this area, it is best to isolate each one and separately study it.

Clustering [14] is a data mining method that has not taken its real part in the works already quoted although, the most important algorithm of this method was very studied in the context of privacy preserving, which is k-means algorithm [20]. Surveying privacy preserving k-means clustering approaches apart from other privacy preserving data mining ones is important due to the use of this algorithm in important other areas, like image and signal processing where the problem of security is strongly posed [9]. Most of works in privacy preserving clustering are developed on the k-means algorithm by applying the model of secure multi- party computation on different data distributions (vertically, horizontally and arbitrary partitioned data). In this work, we build an overview of previous works in privacy preserving k-means clustering algorithm based on secure multi-party computation, we classify them according to data distribution, because it is the only parameter which affects the approach to be undertaken. We present the weaknesses and strengths of each proposed solution, in terms of privacy-preserving and computing cost which are among metrics to evaluate any privacy preserving data mining algorithm [3].

2. K-means Clustering Algorithm

Among the formulations of partitional clustering based on the minimization of an objective function, k-means algorithm [20] is the most widely used and studied. Given a dataset D of n entities (objects, data points, items ...) in real p-dimension space R^p and an integer k. The K-means clustering algorithm partitions the dataset D of entities into k disjoint subsets, called clusters. Each cluster is represented by its center which is the centroid of all entities in that subset.

K-means Algorithm

1. Input: k, the number of centers
2. Selecting the first centers $u_1,\ldots, u_k$
3. Repeat
  For each entity g_i
    a. Calculate distances between g_i and different centers
    b. Find minimal distance
    c. Assign the entity g_i to the nearest center
  End for
  For each cluster j
    a. Calculate the new centers: $v_1,\ldots, v_k$
    b. $u_1,= v_1,\ldots, u_k = v_k$
  End for
4. Until no change between new and old centers
5. Output: entities assignments to the final clusters

The need to preserve privacy in k-means algorithm occurs when it is applied on distributed data over several sites, so-called "parties" and that it wishes to do clustering on the union of their datasets. The aim is to prevent a party to see or deduce the data of another party during the execution of the algorithm. This is achieved by using secure multi-party computation that provides a formal model to preserve privacy of data.

3. Secure Multi-Party Computation

Secure Multi-party Computation (SMC) [11, 12] refers to the general problem of secure computation of a function with distributed inputs between parties. The aim is to protect a honest party against dishonest behaviour of the other party. There are two types of adversaries for which security models are defined, the semi-honest adversary model and the malicious adversary model. The semi-honest model is simpler and it is more answered in privacy preserving applications. It follows the rules of the protocol using its correct inputs, but is free to later use what it sees during execution of the protocol to compromise security. A formal definition of secure (private) two-party computation in the semi honest model is given below: Let f a function defined: $f:\{0,1\}^*\times\{0,1\}^*\rightarrow \{0,1\}^*\times\{0,1\}^*$, where $f=(f_1,f_2)$ Let Π two-party protocol for computing the function f. The view of the first (resp., the second) party during an execution of Π on (x,y) denoted $VIEW^{\Pi}{}_1$ (resp., $VIEW^{\Pi}{}_2$), is: $VIEW^{\Pi}{}_1 =(x, r, m_1,\ldots,m_t)$ and $VIEW^{\Pi}{}_2=(y, r, m_1,\ldots,m_t)$, where r represent the outcome of the first (resp. , second) coin tosses, and m_i represents the i^{th} message it has received. The result of the first (resp., second) party during an execution of Π on (x,y), denoted $OUTPUT^{\Pi}{}_1$ (resp. $OUTPUT^{\Pi}{}_2$) is implicit in the party's own view of the execution, and:

$$OUTPUT^{\Pi}(x,y) = (OUTPUT^{\Pi}{}_1(x,y), OUTPUT^{\Pi}{}_2(x,y)) \quad (1)$$

We say that the protocol Π is secure against semi-honest adversary (compute privately f) if exist probabilistic polynomial-time algorithms, denoted S_1 and S_2, such that:

$$\begin{aligned}\{(S_1(x, f_1(x,y)), f(x,y))\}_{x,y} &\overset{C}{\equiv} \{(VIEW_1^{\pi}(x,y), OUTPUT^{\pi}(x,y))\}_{x,y}\\ \{(S_2(x, f_2(x,y)), f(x,y))\}_{x,y} &\overset{C}{\equiv} \{(VIEW_2^{\pi}(x,y), OUTPUT^{\pi}(x,y))\}_{x,y}\end{aligned} \quad (2)$$

Where $\overset{C}{\equiv}$ denotes computational indistiguishability by (non-uniform) families of polynomial-size circuits.

Informally, the proof of privacy is given by comparing the security in the real world, where the parties use a secure two party protocol, with that in the ideal world, where the parties send their inputs to a trusted party who computes the function and returns only the result to the parties. To ensure privacy in the semi-honest model, security tools are used in the different operations of k-means algorithm. We describe the principal tools.

3.1. Secure Evaluation Circuit

Yao [33] has demonstrated that any polynomially computable function can be computed securely. This was accomplished by demonstrating that given a polynomial size boolean circuit with inputs split between parties, the circuit could be evaluated so that neither side would learn anything but the result. The idea is based on share splitting, the value of each wire in the ciruit is split into two random shares. Secure circuit evaluation does enable efficient computation of functions of small inputs (such as comparing two numbers) because of its prohibitive computation cost, and is used frequently in k-means algorithm for secure comparison distances.

3.2. Homomorphic Schemes

The homomorphism property of these schemes allows to a third party to operate on hidden values, to have a hidden value of the result. Two homomorphic schemes are used in privacy preserving k-means algorithm in particular for secure computation of distances and cluster centers: homomorphic public-key cryptosystems and secret additive sharing schemes. Homomorphic public-key Cryptosystem or homomorphic encryption scheme is a semantically-secure public-key encryption which, in addition to standard guarantees satisfies the following properties:

$$C(t_1) \,.\, C(t_2) = C(t_1+t_2) \quad (3)$$

$$C(t_1)^{t2} = C(t_1 t_2) \quad (4)$$

Where C is the public-key encryption function and t_1, t_2 are elements in the domain of data. The most used are Paillier cryptosystems which provides fast encryption and decryption algorithms [22]. One of its applications is the secure scalar product [10], which is most often used in the k-means algorithm. A secret sharing scheme [27] allows to any t out of n participants to collaboratively recover a secret, while a

set of less than t participants learns nothing about it. *A (t, n)* secret sharing scheme is a set of two functions, the function P which takes a secret s as input and creates n secret shares: $P(s) = (s_1,\ldots,s_n)$ and the reconstruction function R, that return the secret s by gathering the t shares, or nothing. A secret sharing scheme is additively homomorphic if:

$$R(s_1+s'_1, \ldots, s_t+s'_t) = s+s' \quad (5)$$

Pedersen *et al.* [23] show that additive secret sharing schemes are more effective in communication and computation cost compared to the homomorphic cryptosystems when they are applied to privacy preserving data mining. In [6], the authors show an application of additive secret sharing schemes on k-means algorithm.

4. Works in Privacy-Preserving K-Means: Typology and Viewpoints

Several privacy-preserving k-means clustering algorithm are developed on different data distributions. There are two basic data partitioning / data distribution models: horizontal partitioning (homogeneous distribution) and vertical partitioning (heterogeneous distribution).

There are other models of data distribution, such as arbitrarily partitioned data model [13] which generalize the two cases (vertical and horizontal partitioning), where different features for different entities of the same dataset can be owned by several parties. The distribution of datasets across multiple locations dictates changes in the k-means algorithm depending on the model of this distribution, which varies the way to preserve privacy.

4.1. Works in Privacy-Preserving K-Means on Vertically Partitioned Data: Progressions

In a vertically partitioned dataset, each entity is distributed on different parties, so that each party has some components of entities. In this case, the choice of the k first centers does not pose a problem of privacy. While in the iterations of k-means, the computation of the distances over parties involved the need to disclose their data. The intermediate assignments of entities to the nearest clusters also pose a problem of privacy, when searching the minimum of these distances for a given entity. When computing new centers, the old centers shares are explicitly known to the various parties, but the number of entities must be disclosed for the division operation. Hence, the need of privacy comes in distances and centers computation and allocation of entities to the nearest clusters.

The first solution was proposed by Vaidya [31] for vertically partitioned data on several parties. In the proposed protocol, the entities of each party are kept confidential, secure computing distances between the parties is carried out by introducing the secure permutation of Du and Attallah [7] and the homomorphic encryption schemes. The secure comparison between distances is achieved by the Yao [33] evaluation circuit. These primitives are executed for each entity of the dataset which makes the computational cost very high. The protocol requires three non-colluding parties, which have more information than others, such as the partial sum of distances shares, in order to implement the two-party secure comparison and the permutation order of distances vector. The demonstration of privacy is given in the semi-honest model but the protocol reveals additional information such as the assignments of intermediate clusters and the number of entities in each cluster during the operation of division for computing centers.

To avoid the use of non-colluding parties and apply the protocol on two parties, Samet [26] proposes an algorithm to preserve privacy on vertically partitioned data using a new primitive of comparison based on secure multi-party addition developed by the same authors, and the secure sum [5] in computing the sum of distances. However, the protocol does not solve the problem of revelation entities number in clusters. Doganay *et al.* [6] have repeated the same scheme of Vaidya [30] but by using the additive secret sharing schemes instead of homomorphic crypto systems, the main aim is to minimize the computation and communication cost, and to demonstrate the effectiveness of additive secret sharing schemes comparing to the homomorphic crypto systems for these parameters, but by using four non-colluding parties instead of three. An experimental evaluation is also showed.

4.2. Works in Privacy-Preserving K-Means on Horizontally Partitioned Data: Progressions

When k-means algorithm is executed on a horizontally partitioned dataset, the distance computation itself does not violate privacy because each party holds all the components of an entity. The problem arises when computing intermediate cluster centers, in this case, the entities of the same cluster may come from several parties, where the interest of protecting them. This step also requires knowledge of the number of entities in each cluster, this number is extra information that should not be revealed to different parties during the execution of the protocol. The random selection of k first centers is also a problem of privacy in this data distribution. The privacy-preserving protocol on horizontally partitioned data should also prevent the disclosure of additional information such as intermediate centers themselves.

Jha *et al.* [15] have proposed two protocols for the preservation of privacy in k-means algorithm on two

parties only. In the described scenario, only the entities of each party are kept confidential, the intermediate centers are revealed to the two parties and the distance computation is done locally in each party. Security primitives are used for centers computation in order to preserve the privacy of entities of each party. The first protocol (called OPE) is based on oblivious polynomial evaluation given by Naor and Pinkas [21]. The second protocol (named DPE) is based on homomorphic encryption schemes.

The proposed solution still requires that both parties are semi - honest, the aim was to experimentally evaluate the two protocols. The homomorphic encryption scheme is more efficient than OPE for the two parameters: computing and communication cost, but their solution can not be extended to several parties. Samet *et al.* [26] have proposed a protocol which uses a secure method of division that protects the entities of each party and prevents the revelation of the number of entities in each cluster. The protocol is also applicable in a multi-party environment, but the intermediate centers are always revealed.

4.3. Advanced Works in Privacy- Preserving K-Means on Arbitrarily Distributed Data

Although, an arbitrarily partitioned data set is unlikely in practice, the interest of considering such a partitioning is that the protocols in this model can be applied both to horizontally partitioned data and vertically partitioned data. In this case the need of privacy in the k-means algorithm includes all the cases already seen.

The idea of privacy-preserving protocol on arbitrarily partitioned data was introduced by Jagannathan and Wright [13]. Their solution preserves the privacy in the k-means algorithm between two parties. No other confidence party is used. The idea is that all the calculated values in the intermediate steps are split to random values (the principle of random shares). The secure scalar product [10] is used for secure computation of distances, and the Yao circuit evaluation is used for secure comparison and calculation of intermediate centers. These security primitives are executed for each entity of the dataset, which refers the protocol to the same problem of which proposed by Vaidya [30] that of a high computation cost for a large data set.

The solution has a weakness at the updated centers where they consider the division as a multiplication by the inverse, which does not implement correctly the k-means algorithm. Another protocol was given by Su *et al.* [28] where they introduce secure data standardization in order to give greater accuracy to the result of clustering. An improvement in security compared to the first protocol is that the comparison of distances and centers computing are done by the secure scalar product, secure oblivious polynomial evaluation and a secure approximation technique [17]. However, the proposed solution does not solve the problem of the local division, which leaks the number of entities in each cluster to different parties during centers computation and the revelation of the intermediate assignments of entities to clusters.

Bunn and Ostrovsky [4] offer a more efficient solution in privacy for the k-means algorithm on arbitrarily partitioned data. The authors provide a rigorous approach to security in the semi-honest model, their protocol does not reveal the intermediate centers and clusters assignments, they also effectively solve the problem of the local division. They develop secure sub protocols for each operation in the k-means algorithm on the basis of Paillier homomorphic cryptosystems [22] especially for secure multi-party division. The authors propose also a secure protocol for random choice of k initial centers. Despite this, their solution fails to make confidential the number of iterations in the algorithm.

Sakuma and Kobayashi [25] give a new protocol but applied to several parties, called nodes. The aim is to provide a protocol for preserving privacy to simple users rather than organizations or companies. In this case, the protocol is scalable according to the number of parties. The idea is to secure the protocol proposed by Kowalsczyk and Vlassis [18] which calculates the average of the values distributed through P2P networks without transferring data to a central depository for computing centers. At this level, privacy is preserved using Paillier homomorphic crypto systems. For the protection of distances parts, Yao evaluation circuit and the random shares principle are introduced. The calculation is done for each node, which makes the protocol scalable and fault tolerant. But there is no improvement in security.

5. Conclusions

Although the used security model is relatively simple (semi – honest model), the protocols proposed do not provide a complete preservation of privacy in k-means algorithm, the problem is in the iterative nature of the algorithm. Besides, the information to protect are numerous: distances values, intermediate assignments to clusters, number of points in each cluster, intermediate centers, number of iterations and the entities values themselves. The model of the data distribution also affects the way that privacy is preserved.

Arbitrarily partitioning dataset is best suited to seek a more general solution. But most work in this distribution model is applied to only two parties except that in [25]. All work on the vertically partitioned dataset model is given on multiple parties, but considering not colluding trust parties [6, 31], where no security guarantees is given if these parties agreed. The cost of privacy is measured relative to the

computing and communication cost. Protocols based on Yao circuit evaluation [33] are very expensive compared to those based on homomorphic crypto systems, but even the cost of encryption in the latter is not negligible. Additive secret sharing schemes are promising because they present a minimal cost of computation, even if their use requires non-colluding parties [6]. However, the majority of these works apply the Yao protocol [33] and homomorphic cryptosystems. Table 1 summarizes these different solutions and their main objectives.

Our work was to prepare an overview of approaches in privacy-preserving k-means algorithm. We have classified these approaches according to the used distributions datasets, while presenting the weaknesses and strengths of each approach. In contrast to [29] which criticizes the correctness aspect, our analysis is interesting to the privacy preserving one. The interest of our contribution is to highlight the real needs of preserving privacy based on data distribution during the execution of k-means algorithm and to derive the nearest work to solve the problem.

Table 1. Summary of privacy-preserving k-means algorithm works.

Authors	Distribution Model	Parties Number *n*	Security Tools	Main Objective
Vaidya and Clifton, 2003 [31]	Vertical	*n>2*	- Secure permutation [7] - Homomorphic encryption schemes - Yao evaluation circuit [33]	The first privacy-preserving k-means algorithm based on secure multi-party computation
Jha *et al.*, 2005 [15]	Horizontal	*2*	- Oblivious polynomial evaluation [21] - Homomorphic encryption schemes	Comparing the two protocols in term of computation and communication cost
Jagannathan and Wright, 2005 [13]	Arbitrary	*2*	- Random shares - Secure scalar product [10] - Yao evaluation circuit	A secure protocol which can be used on the two data distributions: horizontal and vertical
Samet *et al.*, 2007 [26]	Vertical Horizontal	*n*	- Secure multi-party addition [26] - Secure sum [5]	A multi- party privacy preserving in k-means algorithm
Su *et al.*, 2007 [28]	Arbitrary	*2*	- Secure scalar product - Oblivious polynomial evaluation - Secure approximation technique [17]	Secure data standardisation and security improvement
Bunn and Ostrovsky, 2007 [4]	Arbitrary	*2*	- Paillier cryptosystems [22] - Secure scalar product	Resolving the problem of secure multi-party division and a new protocol for secure randomly selection of *k* first centers
Sakuma and Kobayashi, 2008 [25]	Arbitrary	*n*	- Paillier cryptosystems - Random shares - Yao evaluation circuit	Scalable and fault tolerant protocol
Dogany *et al.*, 2008 [6]	Vertical	*n>3*	- Additive secret sharing schemes [27]	A new protocol based on additive secret sharing schemes instead of homomorphic encryption

References

[1] Aggarwal C. and Yu P., "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," *in Proceedings of Privacy-Preserving Data Mining and Algorithms book*, pp. 11-52, 2008.

[2] Agrawal R. and Srikant R., "Privacy-Preserving Data Mining," *in Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas Texas USA, pp. 439-450, 2000.

[3] Bertino E., Dan L., and Wei J., "A Survey of Quantification of Privacy Preserving Data Mining Algorithms," *in Proceedings of Advances in Database Systems*, pp. 183-205, 2008.

[4] Bunn P. and Ostrovsky R., "Secure Two Party K-Means Clustering," *in Proceedings of the 14th ACM Conference on Computer and Communications Security*, Alexandria, Virginia, USA, pp. 486-497, 2007.

[5] Clifton C., Kantarcioglu M., Vaidya J., Lin X., and Zhu M., "Tools for Privacy Preserving Distributed Data Mining," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp. 28-34, 2002.

[6] Doganay M., Pederson T., Saygin Y., Savas E., and Levi A., "Distributed Privacy Preserving Clustering with Additive Secret Sharing," *in Proceedings of the International Workshop on Privacy and Anonymity in Information Society Table*, Nates, France, pp. 3-11, 2008.

[7] Du W. and Atallah M., "Privacy-Preserving Statistical Analysis," *in Proceedings of 17th Annual Computer Security Applications Conference*, New Orleans Louisiana USA, pp. 102-110, 2001.

[8] El-Sisi A., "Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database," *The International Arab Journal of Information Technology*, vol. 7, no. 2, pp. 152-159, 2010.

[9] Erkin Z., Piva A., Katzenbeisser S., Lagendijk R., Shokrollahi J., Neven G., and Barni M., "Protection and Retrieval of Encrypted Multimedia Content: When Cryptography Meets Signal Processing," *EURASIP Journal of Information Security*, vol. 7, no. 17, pp. 1-20, 2007.

[10] Goethals B., Laur S., Lipmaa H., and Miellkainen T., "On Private Scalar Product Protocol Computation for Privacy-Preserving Data Mining," *in Proceedings of Information Security and Cryptology-ICISC 2004*, pp. 104-120, 2005.

[11] Goldreich O., *Foundations of Cryptography*, Draft of a Chapter on General Protocols, the Press Syndicate of the University of Cambridge, 2003.

[12] Goldreich O., Micali S. and Wigderson A., "How to Play any Mental Game-a Completeness Theorem for Protocols with Honest Majority," *in Proceedings of 19th ACM Symposium of the Theory of Computing*, New York, USA, pp. 218-229, 1987.

[13] Jagannathan G. and Wright R., "Privacy-Preserving Distributed K-Means Clustering over Arbitrarily Partitioned Data," *in Proceedings of the 11th ACM SIGKDD International Conference*, Chicago, USA, pp. 593-599, 2005.

[14] Jain A., Murty M., and Flynn P., "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[15] Jha S., Kruger L., and Mc-Daniel P., "Privacy-Preserving Clustering," *in Proceedings of European Symposium on Research in Computer Security*, pp. 397-417, 2005.

[16] Kantarcioglu M. and Clifton C., "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," *IEEE Transactions on Knowledge and Data Engineering Journal*, vol. 16, no. 9, pp. 1026-1037, 2004.

[17] Kiltz E., Leander G., and Malone-Lee J., "Secure Computation of the Mean and Related Statistics," *in Proceedings of Theory of Cryptograhy Conference*, pp. 238-302, 2005.

[18] Kowalczyk W. and Vlassis N., "Newcast EM," *in Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, Whisteler BC, Canada, pp. 713-720, 2005.

[19] Lindell Y. and Pinkas B., "Privacy-Preserving Data Mining," *in Proceedings of Advances in Cryptography*, pp. 36-53, 2000.

[20] MacQueen J., "Some Methods for Classification and Analysis of Multivariate Observations," *in Proceedings of 5th Berkley Symposium Math. Statistics and Probability*, California, USA, pp. 281-296, 1967.

[21] Naor M. and Pinkas B., "Oblivious Transfer and Polynomial Evaluation," *in Proceedings of the 31st ACM Symposium on Theory of Computing*, Atlanta, USA, pp. 245-254, 1999.

[22] Paillier P., "Public-Key Cryptosystems Ensembled on Composite Degree Residuosity Classes," *in Proceedings of the 17th International Conference on Advances in Cryptography*, Prague Czech Republic, pp. 223-238, 1999.

[23] Pedersen T., Savas E., and Saygin Y., "Secret Sharing vs Encryption-Ensembled Techniques for Privacy-Preserving Data Mining," *in Proceedings of Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Manchester, UK, pp. 646-666, 2007.

[24] Quinlan J., "Introduction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[25] Sakuma J. and Kobayashi S., "Large-Scale K-Means Clustering with User-Centric Privacy Preservation," *in Proceedings of Advances in Knowledge Discovery and Data Mining*, Berlin, pp. 320-322, 2008.

[26] Samet S., Miri A., and Orozco-Barbosa L., "Privacy- Preserving K-Means Clustering in Multi- Party Environment," *in Proceedings of International Conference on Security and Cryptography*, Barcelona, Spain, pp. 523-531, 2007.

[27] Shamir A., "How to Share a Secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612-613, 1979.

[28] Su C., Bao F., Zhou J., Takagi T., and Sakurai K., "Privacy-Preserving Two Party K-Means Clustering Via Secure Approximation," *in Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops*, Oantario, Canada, pp. 385-391, 2007.

[29] Su C., Bao F., Zhou J., Takagi T., and Sakurai K., "Security and Correctness Analysis on Privacy-Preserving K-Means Clustering Shemes," *IEICE TRANSACTIONS on Electronics Communication and Computer Sciences Journal*, vol. 92, no. 4, pp. 1246-1250, 2009.

[30] Vaidya J., "A Survey of Privacy-Preserving Methods across Vertically Partitioned Data," *in Proceedings of Privacy-Preserving Data Mining and Algorithms Book*, pp. 337-358, 2008.

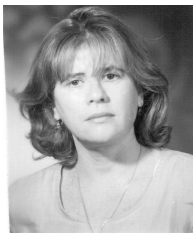
[31] Vaidya J. and Clifton C., "Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data," *in Proceedings of the 9th ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, Washington, USA, pp. 206-215, 2003.

[32] Verykios V., Bertino E., Fovino I., Provenza L., Saygin Y., and Theodoridis Y., "State-of-the-Art in Privacy Preserving Data Mining," In *ACM*

SIGMOD Record, vol. 33, no. 1, pp. 50-57, 2004.

[33] Yao A., "How to Generate and Exchange Secrets," *in Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*, Toronto, Canada, pp. 162-167, 1986.

Fatima Meskine is a doctoral student at the department of computer science, faculty at the University of Oran, and a member of information and information technology laboratory of Oran (LITIO) approved in 2009. She is also the engineer of the computer science center at Mostaganem Univeristy.



Safia Nait Bahloul is a lecturer at the Computer Science Department, Science Faculty at the University of Oran Es- Senia. She obtained, after several scientific stays in CNAM of Paris and LIRIS Laboratory, the University of Claude Bernard 1-Lyon, her PhD degree at the University of Oran in 1997. She is a member of Information and Information Technology Laboratory of Oran (LITIO) which is approved in 2009. Since 2011, she has been leading a team on the Topic of Data Engineering and Web Technology. Her research covers advanced aspects of databases, web technology and unsupervised classification. Her works have been published in several international journals and conferences.