Research Journal of Applied Sciences, Engineering and Technology 10(6): 640-651, 2015
DOI:10.19026/rjaset.10.2473
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: December 29, 2014 Accepted: January 27, 2015 Published: June 20, 2015

Research Article

Privacy Preserving Multiview Point Based BAT Clustering Algorithm and Graph Kernel Method for Data Disambiguation on Horizontally Partitioned Data

[1]J. Anitha and [2]R. Rangarajan
[1]Department of IT, Sri Ramakrishna Engineering College, Vattamalaipalayam,
[2]Department of ECE, RVS College of Engineering and Technology, Coimbatore,
Tamil Nadu 641022, India

Abstract: Data mining has been a popular research area for more than a decade due to its vast spectrum of applications. However, the popularity and wide availability of data mining tools also raised concerns about the privacy of individuals. Thus, the burden of data privacy protection falls on the shoulder of the data holder and data disambiguation problem occurs in the data matrix, anonymized data becomes less secure. All of the existing privacy preservation clustering methods performs clustering based on single point of view, which is the origin, while the latter utilizes many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. To solve this all of above mentioned problems, this study presents a multiview point based clustering methods for anonymized data. Before that data disambiguation problem is solved by using Ramon-Gartner Subtree Graph Kernel (RGSGK), where the weight values are assigned and kernel value is determined for disambiguated data. Obtain privacy by anonymization, where the data is encrypted with secure key is obtained by the Ring-Based Fully Homomorphic Encryption (RBFHE). In order to group the anonymize data, in this study BAT clustering method is proposed based on multiview point based similarity measurement and the proposed method is called as MVBAT. However in this paper initially distance matrix is calculated and using which similarity matrix and dissimilarity matrix is formed. The experimental result of the proposed MVBAT Clustering algorithm is compared with conventional methods in terms of the F-Measure, running time, privacy loss and utility loss. RBFHE encryption results is also compared with existing methods in terms of the communication cost for UCI machine learning datasets such as adult dataset and house dataset.

Keywords: BAT algorithm, cluster analysis, data disambiguation, data mining, distributed multi view point based clustering, graph partitioning, horizontal partitioning data, privacy, Ramon-Gartner Subtree Graph Kernel (RGSGK), Ring-Based Fully Homomorphic Encryption (RBFHE), security

INTRODUCTION

Data mining is used to extract implicit and previously unknown information from data. Data mining is the process which provides a concept to attract attention of users due to high availability of huge amount of data and need to convert such data into useful information. Naturally this raised privacy concerns about collected data. In response to that, data mining researchers started to address privacy concerns by developing special data mining techniques under the framework of "privacy preserving data mining". Opposed to regular data mining techniques, privacy preserving data mining can be applied to databases without violating the privacy of individuals. Privacy preserving techniques for many data mining models have been proposed in the past 5 years. Techniques for privacy preserving association rule mining in distributed environments (Kantarcioglu and Clifton, 2004).

The privacy violation through the process of mining can pose real privacy issues. The reason is that gathering data and bringing them together to support data mining makes misuse easier. In other words, the problem is not data mining results, but the process that generates them. If the results were generated without sharing information and the results could not be used to deduce private information, data mining would not reduce privacy (Vaidya and Clifton, 2003). Although obtaining globally meaningful results without sharing information seems impossible, some solutions have been proposed for that. In order to perform privacy preservation concept some of the methods have been proposed in earlier work for horizontally partitioned data. ID3 classification (Lindell and Pinkas, 2000) for two parties with horizontally partitioned data by using

Corresponding Author: J. Anitha, Department of IT, Sri Ramakrishna Engineering College, Vattamalaipalayam, Coimbatore, Tamil Nadu 641022, India

secure protocols to achieve complete zero knowledge leakage. Four efficient methods (Clifton *et al*., 2003) namely secure sum, secure set union, secure size of set intersection and scalar product for privacy preserving data mining in distributed environment. Privacy preserving data mining of association rules (Kantarcioglu and Clifton, 2004) when the data is partitioned horizontally. They proposed algorithm which uses three basic ideas such as randomization, encryption of site results and secure computation. The state of art in the area of privacy preserving data mining techniques is presented (Verykios *et al*., 2004). The authors also discussed about classifications of privacy preserving techniques and privacy preserving algorithms such as heuristic-based techniques, cryptography-based techniques and reconstruction based technique. A framework for evaluating privacy preserving data mining algorithms and based on this frame work one can assess the different features of privacy preserving algorithms according to different evaluation criteria (Elisa *et al*., 2005).

Clustering is widely used in many applications such as customer behavior analysis, targeted marketing and others. Recently, privacy preserving clustering problems has also been studied by many authors. Existing privacy-preserving protocols based on the k-means algorithm, Fuzzy c means clustering and this protocol does not reveal intermediate candidate cluster centers. These existing solutions can be made more secure but only at the cost of a high communication complexity. Data containers need to send their data to the third party and at the same time they need to keep privacy on data not solved by existing work, all of the existing work doesn't perform multi view point based clustering for anonymized data, data disambiguation problems is also not solved by these methods. Simultaneously, clustering still requires more robust dissimilarity or similarity measures; recent works such as (Lee and Lee, 2010) illustrate this need.

Similarity measure plays a very important role in the success or failure of a clustering method. Our first objective is to derive a novel method for measuring similarity between data objects in sparse and high-dimensional domain, particularly anonymized data. From the proposed similarity measure, then formulate new clustering criterion functions. The objective of our work is to develop a privacy preserving multiview point based clustering method for horizontally partitioned data on only two parties. In distributed architecture, the numbers of data containers are connected with the single third party that knows the multi view clustering procedure. Before performing the multi view point data clustering for horizontally partitioned data the data disambiguation and anonymization problems is solved for data holder. The data disambiguation problems are solved by using RGSGK. Then anonymize the data by encrypting the original data with the secure key before clustering the data, in order to achieve privacy by using MDHKEA. To allow parties to obtain the final results without revealing intermediate candidate cluster centers, propose RBFHE methods for secure computation. The proposed study privacy preserving MVBAT-means clustering for horizontally partitioned data is performed between two parties. Thus, parties cannot learn extra information of the others.

LITERATURE REVIEW

DBSCAN (Liu *et al*., 2012) is a well-known density-based clustering algorithm which offers advantages for finding clusters of arbitrary shapes compared to partitioning and hierarchical clustering methods. However, there are few papers studying the DBSCAN algorithm under the privacy preserving distributed data mining model, in which the data is distributed between two or more parties and the parties cooperate to obtain the clustering results without revealing the data at the individual parties. Address the problem of two-party privacy preserving DBSCAN clustering. First propose two protocols for privacy preserving DBSCAN clustering over horizontally and vertically partitioned data respectively and then extend them to arbitrarily partitioned data.

İnan *et al*. (2007) propose methods for constructing the dissimilarity matrix of objects from different sites in a privacy preserving manner which can be used for privacy preserving clustering as well as database joins, record linkage and other operations that require pair-wise comparison of individual private data objects horizontally distributed to multiple sites. It show communication and computation complexity of our protocol by conducting experiments over synthetically generated and real datasets.

Privacy-preserving collaborative filtering algorithm (Jeckmans *et al*., 2012), which allows one company to generate recommendations, based on its own customer data and the customer data from other companies. The security property is based on rigorous cryptographic techniques and guarantees that no company will leak its customer data to others. In practice, such a guarantee not only protects companies' business incentives but also makes the operation compliant with privacy regulations.

Mangasarian (2012) propose a simple privacy-preserving reformulation of a linear program whose equality constraint matrix is partitioned into groups of rows. Each group of matrix rows and its corresponding right hand side vector are owned by a distinct private entity that is unwilling to share or make public its row group or right hand side vector. By multiplying each privately held constraint group by an appropriately generated and privately held random matrix, the original linear program is transformed into an equivalent one that does not reveal any of the privately held data or make it public. The solution vector of the

transformed secure linear program is publicly generated and is available to all entities.

Two-Party k-Means Clustering Protocol (Bunn and Ostrovsky, 2007) that guarantees privacy and is more efficient than utilizing a general multiparty "compiler" to achieve the same task. In particular, a main contribution of our result is a way to compute efficiently multiple iterations of k-means clustering without revealing the intermediate values. To achieve this, use novel techniques to perform two-party division and sample uniformly at random from an unknown domain size.

Privacy preserving hierarchical k-means clustering algorithm on horizontally partitioned data, denoted as HPPHKC (Xue *et al*., 2009). The algorithm has two phases: the first phase, every object can be as a cluster, a secure computation protocol is used to compute the dissimilarity matrix and the most similar clusters will be merged. This process is repeated until get the assigned clusters number k and get k clustering centers. In the second phase, the semi-honest third party and all data involved parties use the k-means algorithm refine the results of the first phase and get the final clustering results.

All of the above clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. Traditional dissimilarity/similarity measure perform clustering single viewpoint, it reduces the clustering accuracy for anonymized data where assumed to be in the same cluster with the two objects being measured. To overcome this above mentioned problem, proposed work using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim.

PROPOSED METHODOLOGY

In this study a novel horizontal partitioning approach for multiview point clustering anonymized data is proposed. Before anonymization is performed for multiview point clustered data it becomes important to secure the data and hence anonymize the original data with the help of encryption technique. The secure key fulfils the encryption process in order to achieve the secure key; where use the key generation algorithm namely Ring-Based Fully Homomorphic Encryption (RBFHE). In this study data ambiguation problem occurs by blanking certain fields in the data table in such a way that no entry (row) in the table is unique. This makes it impossible to uniquely identify an entry by linking to another data table, since in an ambiguated table; at least two rows will match any linking operation. The same fields occurs in the table also occurs for the table it also critical to solve data disambiguation problem. After anonymize the data it is grouped based on single view point that is measuring similarity between the inter

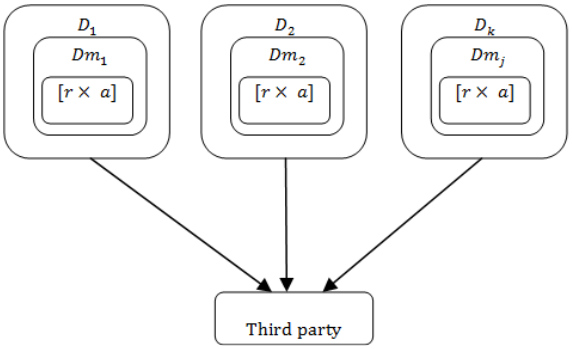

Fig. 1: Illustrates the distributed architecture of data holder and third party

cluster similarity and dissimilarity wise only. But measuring the intra cluster similarity based measurement also important to perform clustering process. In order to solve these problem formally convert the data table into Ramon-Gartner subtree graph kernel method, then it finds the repeated attributes that have occurs the same attribute value in the table. This data disambiguation problem is solved by using Ramon-Gartner subtree kernel. Then multiview point based similarity measurement is performed for data points of the anonymized data. In this paper, attain privacy of cluster by the following three steps:

- Solve data disambiguation problem by the Ramon-Gartner Subtree Graph Kernel (RGSGK) method.
- Anonymize the original data with the secure key, by using Ring-Based Fully Homomorphic Encryption (RBFHE).
- Multiview point based BAT cluster algorithm for anonymize data it is named as MVBAT Clustering. The proposed clustering methods is used to cluster the anonymize data in multiview point manner.

In order to perform this process first need to formally define the problem; give details on trust levels of the involved parties and the amount of preliminary information that must be known by each one. There are k data holders, such that $k \geq 2$, each of which owns a horizontal partition of the data matrix D, denoted as D_k. It consists of k data holders D_k and single third party TP. The each data holder D_k consists of data matrix Dm_j and the data matrix consists of the a number of attributes and, b number of objects $[r \times a]$. The distributed architecture is given in Fig. 1.

Ramon-Gartner Subtree Graph Kernel (RGSGK) method for data disambiguation: Table 1 the illustrates the data matrix of the each data holder, where the disambiguation data presents and after disambiguated data is found in the table by RGSGK method then convert those data into order manner by highest attribute value. In Table 1, let us consider $G = G(V, E, L)$ be an undirected graph where V is a set

Table 1: Illustrates the data matrix of the data holder

ID	Age	Sex	BP	Cholesterol mg/dL	Sugar	Heart rate	Heart patient
1	70	1	130	322	1	109	1
2	67	0	115	564	0	160	2
3	70	1	130	324	1	109	1
4	64	1	128	263	0	105	2
5	70	1	130	325	1	109	1
6	56	0	130	256	1	142	2
7	57	0	128	254	1	141	2

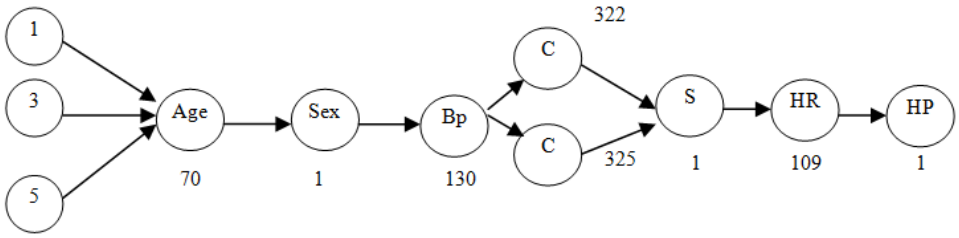

Fig. 2: Example of the graph

of vertices and E a set of edges.Each attributes in the data matrix is represented as a vertex $v \in V$ in the graph and an edge $e \in E$ is added to the graph for every pair of vertices representing attributes which can potentially be the same heart patient which belongs to either one or two, L be the label of the graph kernel, that is the vertices assigns the labels names to nodes such as Age, Sex, Blood Pressure (BP), Cholesterol, Sugar, Heart rate and Heart patient. The heart patient records which belongs to one is represented as same graph and heart which belongs to category two represented as another graph. In order to perform the data disambiguation problem the set of the following constraints are represented between the different data attributes in the graph (Bach, 2008).

The neighbourhood $N(v)$ of a node v is the set of nodes to which v is connected by an edge, that is $N(v) = \{v' | (v; v') \in E\}$. For simplicity, assume that every graph has n nodes, m edges, a maximum degree of d and that there are N graphs in our given set of graphs. Let S (G) refer the set of subtree patterns in the graph in this study there a two types of the graphs based on the heart patient type. The first subtree kernel on the graph was defined by Ramon and Gartner (2003). It compares the pairs of nodes from different patterns in the graphs $p_1 = (V, E, L), p_2 = (V', E', L')V$ represents vertices of the graph of the current data sampleand V' represents vertices of the graph (Fig. 2) of the remaining samples which are represented in the graph by iteratively comparing their neighbourhoods:

$$k_{ramon}^{(h)}(p_1, p_2) = \sum_{v \in V, v' \in V'} k_h(v, v') \quad (1)$$

$$k_h(v, v') = \begin{cases} \alpha * \delta(L(v)L(v)), \text{if } h = 1 \\ \lambda_k a_k(x_n, x_m, r_{nm}), \text{if } h > 1 \end{cases} \quad (2)$$

α be the randomly assigning weight values for same attribute value, for same attributes with same value the h = 1. If the attributes values are different then h is greater than one. The set of relations between two records (x_n and x_m), even direct or indirect, are represented as r_{nm}. λ_k is the weight applied to the attribute value if both belongs to same attribute with different values. In our table convert this into graph in the following manner.

Here applied only the variable weight value for two different rows which have maintain the same attribute values and it belongs to h = 1, assign α weight value if it is higher value than all the remaining records higher weight value is also assigned to this attribute value for same heart patients. So disambiguation problem is applicable to rows 1, 3, 5 in the Table 1. In this example the record 1, 3 have all the values of the attributes have same values, so disambiguation occurs. Final weight value is applied to the record 1 as:

$$k_h(70,70) = 0.9 \times 70 = 63 \quad (3)$$

$$k_h(67,64) = 0.85 \times 67 = 57 \quad (4)$$

Similarly it is also applied to entire record 1 in the table then values are converted based on this calculated value, similarly it is also applied to record 3 for all attributes since it comes under as the second part in the table, so the disambiguation problem is solved by calculation of the weight values and then original table values also changed it improves the privacy accuracy since original value are changed as unknown values. Based the attribute value the weight is multiplied as high for individual data in the Table 1. RGSGK task determines the best disambiguation problem results for the vertices, given a set of conditions. In this case, the conditions are the edge weights, which represent how strong are the involved constraints. Positive weights indicate that both adjacent vertices should be in the same attribute value in the records from Table 1. Data

Table 2: Illustrates the data matrix of the data holder after RGSGK

ID	Age	Sex	BP	Cholesterol mg/dL	Sugar	Heart rate	Heart patient
1	63	1	117.00	254.38	1	92.65	1
2	57	0	92.00	507.60	0	152.00	2
3	56	1	115.70	259.20	1	91.56	1
4	51	1	107.52	205.14	0	86.10	2
5	49	1	114.40	263.25	1	90.47	1
6	42	0	111.80	191.12	1	127.80	2
7	41	0	106.24	193.60	1	127.50	2

disambiguation problem solved table results is shown in Table 2.

Each data holder needs to cluster their data, the cluster algorithm is available in the third party but the third party and the other data containers are semi trusted. So, if the data is send directly to the third party then whole data may be known by, all other data holder and third party. Since, there is a necessity to anonymize the original data before sending the data to the third party. Here, apply the RBFHE encryption process to the entire original data with secure key to anonymize the original data. That secret key is important aspect for achieving the privacy of data. With the help of the RBFHE. keygen $(d, q, t, \chi_{key}, \chi_{err}, w)$ algorithm, attain the secure key. Third party's duty in the protocol is to govern the communication between data holders, construct the dissimilarity matrix and publish clustering results to data holders.

Ring-Based Fully Homomorphic Encryption (RBFHE) for data anonymization: The proposed RBFHE construction of key is developed in third party and in data holders since both is semi trusted. The third party generates one public key for all data holder that will send the network publicly with hiding some important value. Every data holder calculates a new private key by the received public value from the third party. The generation of the secure key has two steps mainly they are:

- Public key generation in third party
- Secret key generation in data holder

The entire procedure for proposed RBFHE encryption schema for anonymize the data is specified in detail in the following way. In this ciphertext consists of only a single ring element as opposed to the two or more ring elements for schemes based purely on the (ring) learning with errors. The scheme is scale-invariant and therefore avoids modulus switching and the size of ciphertexts is one ring element. The data entry of each data holder samples dms_i and $\varphi(rn, hk)$ is the key value for each data holder. The most important structure is the ring R. To perform the encryption and decryption process for anonymize data define some the parameters, Let d be a positive integer and define:

$$R = Z[dms_i]/(\phi_d(dms_i)) \quad (5)$$

As the ring of polynomials with integer coefficients modulo the d-th cyclotomic polynomial $\phi_d(dms_i) \in Z[dms_i]$. The degree of ϕ_d is $rn = \varphi(d)$ where φ is Euler's totient function for data accountability for each data holder data matrix samples dms_i. The elements of R that is each data holder data matrix samplesdms_i can be uniquely represented by all polynomials in $Z[dms_i]$ of degree less than rn. Arithmetic in R is arithmetic modulo $\phi_d(dms_i)$ which is implicit whenever write down terms or equalities involving elements in R. The arbitary coefficient that belongs to the each data holder data matrix samples dms_i in R:

$$a = \sum_{i=0}^{n-1} a_i\, dms_i \in R||a||_\infty = \max_i\{|a_i|\} \quad (6)$$

where, $a_i \in Z$ identify a with its vector of coefficients of the attributes in the data holder data matrix samples and choose maximum data holder data matrix samples with $\mathbb{R}^n$ to measure the size of elements in R. When multiplying two elements $g, hk \in R$, the norm of their product g, hk expands with respect to the individual norms of g and hk. The maximal norm expansion that can occur:

$$\delta = \sup\left\{1\left|\frac{|g.h||_\infty}{||g||_\infty||h||_\infty}\right.\right\}; g, hk \in R \quad (7)$$

Which $g, hk \in R$ is a ring constant. Let χ be a probability distribution on R that samples small elements $a \leftarrow \chi$ with high probability. The distribution χ on R is called B-bounded for some $B > 0$ if for all $a \leftarrow \chi$ and have $\left|||h|\right|_\infty < B$. First, define the discrete Gaussian distribution $D_{z,\sigma}$ with mean 0 and standard deviation σ over the integers, which assigns a probability proportional to $\exp(-\pi |r|^2/\sigma^2)$ to each data holder data matrix samples $dms_i \in Z$ and when d is a power of 2 and $\phi_d(r) = dms_i^{rn} + 1$ take χ be a spherical discrete gaussian probability distribution $\chi = D_{Z^n,\sigma}$ where each coefficient of dms_i is sampled according to the one dimensional distribution. The distribution is used in many fully homomorphic encryption schemes based on RBFHE with high probability.

The public key encryption scheme is parameterized by a modulus q and a plaintext modulus $1 < t < q$. The secret key S of each data holder attribute value is derived from the distribution key χ_{key} and errors are sampled from the distribution χ_{err}. The basic encryption and decryption steps of the data holder data matrix samples are defined as below.

Basic: Params Gen (λ): Given the security parameter λ, a positive integer d that determines R, modulo q and t with $1 < t < q$ and distributions χ_{key}, χ_{err} on R output of decrypted data $(d, q, t, \chi_{key}, \chi_{err})$.

Basic keyGen $(d, q. t, \chi_{key}, \chi_{err})$ data holder data matrix samples $dms_i^{'} g \leftarrow \chi_{key}$ and let $dms_i = [tdms_i^{'} + 1]_q$. If dms_i is not invertible to modulo q choose new $dms_i^{'}$. Compute inverse $dms_i^{-1} \in R$ of dms_i modulo q and set:

$$h = [tgdms_i^{-1}\varphi(rn, hk)]_q \quad (8)$$

Output the public and private key pair:

$$(pk, sk) = (hk, dms_i, \varphi(rn, hk)) \in R^2 \quad (9)$$

Basic: Encrypt (hk, m): the message space is Rt/R for message $m + tR + \varphi(rn, hk)$ choose $[m]_t$ as its representative. Sample s, $e \leftarrow \chi_{err}$ and output the ciphertext of encrypted data:

$$c = [[q/t][m]_t + e + hs + \varphi(rn, hk)]_q \in R \quad (10)$$

The message or information is defined as:

$$m = [[t/q \,.\, [c]_q]]_t \in R \quad (11)$$

The following lemma states conditions for a ciphertext c such that the decryption algorithm outputs the message m that was originally encrypted data. Given $m \in R$, a public key:

$$h = [dms_i^{-1}\varphi(rn, hk)]_q \quad (12)$$

With secret key:

$$SK = [1 + tdms_i^{'}]_q, dms_i^{'}, g \leftarrow \chi_{key} \quad (13)$$

And let $c = Basic.Encrypt(h, m)$. RBFHE. Parametergen (λ): Given the security parameter λ output of the parameter with encrypted data holder data matrix samples using $(d, q, t, \chi_{key}, \chi_{err}, w)$ where $(d, q, t, \chi_{key}, \chi_{err}) \leftarrow$ BasicParamgen (λ) and $w > 1$ is integer RBFHE. keygen $(d, q, t, \chi_{key}, \chi_{err}, w)$ compute $kg \leftarrow Basic.Keygen(d, q, t, \chi_{key}, \chi_{err})$ sample $e, s \leftarrow \chi_{err}^{\ell_{w,q}^3}$ compute:

$$\gamma = [kg^{-1}P_{w,q}\left(D_{w,q}(kg) \otimes D_{w,q}(kg)\right) + e + h.s]_q \in R^{\ell_{w,q}^3} \quad (14)$$

RBFHE. encrypt $(pk, sk, evk) = (h, kg, \gamma)$
RBFHE. encrypt (pk, m) to encrypt $m \in R$
RBFHE. Decrypt (sk, c) to output the message encrypt $m \leftarrow Basic.Decrypt(sk, c) \in R$
RBFHE. KeySwitch $(\tilde{c}_{multi}), evk)$: output $[\langle D_{w,q}(\tilde{c}_{multi}), evk\rangle]_q \in R$.
RBFHE. add (C_1, C_2): Compute the addition of C_1, C_2 as $C_{add} = [C_1 + C_2]_q$
RBFHE. multi (C_1, C_2, evk)

Compute:

$$\tilde{c}_{multi} = \left[\left[\frac{t}{q}P_{w,q}(C_1) \otimes P_{w,q}(C_2)\right]\right]_q \in R^{t_{w,q}^2} \quad (15)$$

And output $c_{multi} = RBFHE.multi(\tilde{c}_{multi}, evk)$.

Given two ciphertexts $C_1, C_2 \in R$ which encrypt two messages m_1, m_2 with inherent noise terms v_1, v_2 their sum modulo q, $C_{add} = [C_1 + C_2]_q$ encrypts the sum of the message modulo t $[m_1 + m_2]_t$ and rewrite this as $[m_1 + m_2]_t + tr_{add} = [m_1]_t + [m_2]_t$ for some $dms_{i_{add}} \in R$ with $||dms_{i_{add}}||_\infty \leq 1$:

$$kg[C_1 + C_2]_q = kgC_1 + kgC_2 = \Delta([m_1]_t + [m_2]_t] + (v_1 + v_2) = \Delta(\left[m_1 + m_2\right]_t + tdms_{i_{add}} + (v_1 + v2 \ (mod\ q) \quad (16)$$

This means that the size of the inherent noise v_{add} of c_{add} is bounded by:

$$|v_{add}||_\infty \leq ||v_1||_\infty + ||v_2||_\infty + dms_{i_t}(q) \quad (17)$$

Homomorphic Multiplication operation is divided into two parts. The first part describes a basic procedure to obtain an intermediate ciphertext that encrypts the product $[m_1 m_2]_t$ modulo t of two messages m_1 and m_2. The second part performs a procedure which allows a public transformation of this intermediate ciphertext to a ciphertext that can be decrypted. This latter procedure was introduced (Brakerski and Vaikuntanathan, 2011) in the form of relinearization and was later expanded (Brakerski *et al*., 2012) into a method called key switching, which transforms a ciphertext decryptable under one secret key to one decryptable under any other secret key. For our analysis, assume that χ_{key}, χ_{err} respectively. RBFHE. multi (C_1, C_2, evk) compute:

$$\tilde{c}_{multi} = \left[\left[\frac{t}{q}P_{w,q}(C_1) \otimes P_{w,q}(C_2)\right]\right]_q \in R^{t_{w,q}^2} \quad (18)$$

The second part in the homomorphic multiplication procedure is a key switching step, which transforms the ciphertext $\tilde{c}_{multi}$ into a ciphertext C that is decryptable under the original secret key:

$$evk = \left[kg^{-1}P_{w,q}\left(D_{w,q}(kg) \otimes D_{w,q}(kg)\right) + e + hk.s\right]_q \quad (19)$$

Output by RBFHE. Keygen where $e, s \leftarrow \chi_{err}^{-w,q}$ are vectors of polynomials sampled from the error distribution χ_{err} and $[.]_q$ is applied to each coefficient of the vector and that it is made public because it is needed for the homomorphic multiplication operation. Every data holder and the third party must have access to the comparison functions so that they can compute distance/dissimilarity between objects for clustering the anonymize data. Data holders are supposed to have agreed on the list of attributes that are going to be used for clustering beforehand. This attribute list is also shared with the third party so that TP can run appropriate comparison functions for different data types. At the end of the protocol, the third party will have constructed the dissimilarity matrices for each attribute separately.

Multiview point based BAT clustering for anonymized data: Third party only gets the cipher text from all the data holders in the network. To perform MVBAT Clustering methods for cipher text of all data holders idealize some of the echolocation characteristics of microbats, can develop bat algorithms. For simplicity, now use the following approximate rules to perform multiview point based clustering method:

- All ciphertext of the data holder bats use echolocation to sense distance and they also 'know' the difference between food/prey and background barriers in some magical way.
- Bats fly randomly with velocity v_i at position $\tilde{cx}_{multii}$ with a fixed frequency f_{min} varying wavelength λ and loudness A_0 to search for prey. They can automatically adjust the wavelength of their emitted pulses and adjust the rate of pulse emission $r \in [0, 1]$, depending on the proximity of their target; In general the frequency f in a range $[f_{min}, f_{max}] = [20kHz, 500kHz]$ corresponds to a range of wavelengths $[\lambda_{min}, \lambda_{max}] = [0.7mm, 17mm]$. The proximity target function is determined based on the multiview point based clustering method criteria function, in this study two criteria functions $I_o \& I_J$ are used to measure the similarity between two bats cipertext.
- Although the loudness can vary in many ways, assume that the loudness varies from a large (positive) A_0 to a minimum constant value A_{min}.

Movement of virtual bats: In this study use a virtual bats randomly to perform the multiview point based clustering for cipher text data. To perform this process the ciphertext data samples position $x_i = (\tilde{cx}_{multi_1}, \ldots . \tilde{cx}_{multi_n})$ and velocity v_i are updated. The new solutions of the clustered data are represented as x_i^{bt} and velocity v_i^{bt} at specific bat time interval bt are given by:

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad (20)$$

$$v_i^{bt} = v_i^{bt-1} + (x_i^{bt} - x_*)f_i \quad (21)$$

$$x_i^{bt} = x_i^{bt-1} + v_i^{bt} \quad (22)$$

where, $\beta \in [0,1]$ is a random vector drawn from uniform distribution. Here x_* is the current best multiview point based clusterresult which is located after comparing all the solutions among bats. As the product $\lambda_i f_i$ is the velocity increment, use either f_i or λ_i, depending on the type of the problem interest. For local clustering process of multiview point based clustering for anoymized data, once best cluster is found a new solution for each bat is generated locally using random walk:

$$x_{new} = x_{old} + \epsilon A^{bt} \quad (23)$$

where, $\epsilon \in [-1,1]$ is random number while $A^t = < A_i^{bt} >$ is the average loudness of all the bats at this time step.

Loudness and pulse emission: Furthermore, the loudness A_i and the rate r_i of pulse emission have to be updated accordingly two criteria functions $I_o \& I_J$ are used to measure the similarity between two bats cipertext as the iterations proceed. As pulse emission increases it becomes more similarity value to form a cluster for anonymized data, the loudness can be chosen as any value of convenience. For simplicity, can also use $A_0 = 1$ and $A_{min} = 0$:

$$A_i^{bt+1} = \rho A_i^{bt} \quad (24)$$

$$r_i^{bt+1} = r_i^0[1 - \exp(-\gamma bt)] \quad (25)$$

where, ρ and γ are constants. In fact, ρ is similar to the cooling factor, for any $0 < \rho < 1$ and $\gamma > 0$:

$$A_i^{bt+1} \to 0, r_i^{bt+1} \to r_i^0 \text{ as } bt \to \infty \quad (26)$$

In the simplicity case, can use $\rho = \gamma$ and have used $\rho = \gamma = 0.9$ in our simulations. Initial emission rate r_i^0 can be determined using two criteria functions $I_o \& I_J$ based on the calculation of the distance matrix and dissimilarity matrix.

Distance matrix: The distance matrix is used to find the distance between all data points with selected cluster centroids. This distance matrix helps the third party in calculating the similarity matrix and dissimilarity matrix in an easy way. With the help of the distance matrix, can find the similarity matrix. The similarity matrix is $[n * cn]$ matrix where n is number of data points and, cn is the selected cluster centroid. The matrix consists of similarity value of each data point that moves to the cluster centroid. The following

Eq. (27) is used to find the similarity value of data points with each cluster:

$$S_{xy} = \frac{(d_{ij})}{\sum_{j=1}^{n}(d_{ij})} \quad (27)$$

From the above equation S_{xy} the x denotes the data point and y denotes the cluster centroid. Based on the distance from the cluster centroid to the data points, third party calculates the similarity value of each data point. The similarity value of the data point declares, how much the data point is closer with correspond cluster centroid. The data point moves to the cluster centroid which, has the highest similarity value among them.

Dissimilarity matrix: The dissimilarity matrix is also $[n * cn]$ matrix which consists of dissimilarity value of the data point with the cluster centroid. The dissimilarity value describes how much distance is required to the data point go away from the cluster centroid. The following Eq. (28) is used to find the dissimilarity value of data points with each cluster:

$$D_{ij} = \max(d)\left[CT_j\right] - d_{ij} \quad (28)$$

In the above equation i corresponds, to the data point and j corresponds to the cluster centroid. Now have to find the maximum distance of each cluster and subtract with the data point. This result is dissimilarity value of the data point. With the help of the dissimilarity value, the third party can calculates dissimilarity matrix by the following Eq. (29):

$$Dss_{ij} = \frac{D_{ij}}{\sum_{j=1}^{n}(D_{ij})} \quad (29)$$

Their loudness and emission rates will be updated only if the new solutions are improved, which means that these bats are moving towards the optimal solution. The final form of our criterion function I_o is:

$$I_o = \sum_{r=1}^{k} \frac{1}{b_r^{1-\omega}} \left[\frac{b+b_r}{b-b_r}\left|\left|S_{xy}\right|\right|^2 - \left(\frac{b+b_r}{b-b_r} - 1\right) Dss_{ij_r}\right] \quad (30)$$

Dss_{ij_r} denotes the dissimilarity matrix value based on the heart rate value and S_{xy} represents the similarity matrix, b denotes the bat (cipher text value of the data holders), b_r denotes the ciphertext value of the data holder alognwith the Heart patient type. The second criteria function I_J to perform the clustering process is defined as follows:

$$I_J = \sum_{r=1}^{k}\left[\frac{b+\left|\left|S_{xy}\right|\right|}{b-b_r}\left|\left|S_{xy}\right|\right| - \left(\frac{b+\left|\left|S_{xy}\right|\right|}{b-b_r} - 1\right)\frac{Dss_{ij_r}}{\left|\left|S_{xy}\right|\right|}\right] \quad (31)$$

Proposed MVBAT clustering:

Objective function $f(x), x = (x_1, \ldots, x_d)$

Initialize the bat population $x_i\ (i = 1, 2, \ldots, n)$ and vi assigning values from a data matrix from RBFHE

Define pulse frequency f_i at x_i

Initialize pulse rates r_i two criteria functions I_o&I_J form (30) and (31) and the loudness A_i

while ($t < Max\ number\ of\ iterations$)

Generate new solutions by adjusting frequency and updating velocities and locations/solutions Eq. (20) to (23)

if ($rand > r_i$)

Select a best data points solution

Generate a local data point's solution around the selected best data points for multiview point based clustering of anonymized data

end if

Generate a new data point solution by flying randomly

if ($rand < A_i$& $f(x_i) < f(x_*)$), $f(x_i), f(x_*)$ is also obtained based on the two criteria functions I_o&I_J form (30) and (31)

Accept the new data points as cluster points for multiview point based clustering

Increase r_i and reduce A_i

end if

Rank the bats and find the current best x_* data points

end while

Post process results and visualization

Generate a new data point solution by flying randomly

if ($rand < A_i$& $f(x_i) < f(x_*)$), $f(x_i), f(x_*)$ is also obtained based on the two criteria functions I_o& I_J form (30) and (31)

EXPERIMENTAL RESULTS

The experiments for evaluating the performance of proposed MVBAT Clustering for horizontal partitioning data are explained and discussed in detail. Our proposed MVBAT Clustering method for horizontal partitioning data is different from existing clustering methods since the proposed MVBAT Clustering, clustering is performed for anonymized data based on the multiview point, but earlier work focus on single view point based clustering, so it produces less information loss when compare to conventional methods and each attribute value is encrypted by a RBFHE. In order to measure clustering results therefore, perform the following performance evaluation metrics such as communication cost analysis and running time analysis, Information loss, utility, privacy loss and clustering methods accuracy. In our experimentation have used two data sets from UCI Machine Learning Repository (Frank and Asuncion, 2010) datasets such as Adult Dataset and Housing Data

Table 3: Description of the adult data set

Number	Attribute	Type	# of values
1	Age	Continuous	74
2	Work class	Categorical	8
3	Final weight	Continuous	NA
4	Education	Categorical	16
5	Education-num	Continuous	16
6	Martial-status	Categorical	7
7	Occupation	Categorical	14
8	Relationship	Categorical	6
9	Race	Categorical	5
10	Sex	Categorical	2
11	Capital-gain	Continuous	NA
12	Capital-loss	Continuous	NA
13	Hours-per-week	Continuous	NA
14	Country	Categorical	41
15	Salary	Categorical	2

Table 4: Description of the housing data set

Number	Attribute	Type
1	Per capita crime rate by town	Continuous
2	Proportion of residential land zoned	Continuous
3	Proportion of non-retail business acres per town	Continuous
4	Charles river dummy variable	Continuous
5	Nitric oxides concentration	Continuous
6	Average number of rooms per dwelling	Continuous
7	Proportion of owner-occupied units built prior to 1940	Continuous
8	Weighted distances to five boston employment centres	Continuous
9	Index of accessibility to radial highways	Categorical
10	Full-value property-tax rate	Continuous
11	Pupil-teacher ratio by town	Continuous
12	Proportion of blacks by town	Continuous
13	Lower status of the population	Continuous
14	Median value of owner-occupied homes	Continuous

Set are taken here to measure the performance of the proposed GFA clustering method for horizontally partitioning data. Datasets are more appropriate for our experiments since try to evaluate scalability and efficiency of our clustering methods for horizontally partitioned data by varying parameters. Data generator is developed in Eclipse Java environment. Adult data set from the UC Irvine machine learning repository which is comprised of data collected from the US census. The data set is described in Table 3. Tuples with missing values are eliminated and there are 45, 222 valid tuples in total. The adult data set contains 15 attributes in total.

The Housing Data Set is described in Table 4. It totally contains 14 attributes in total. Divide datasets into datasets of size 2, 4, 6, 8 and 10 K, respectively where K represents thousands.

In our experiments, use RBFHE cipher to generate private key for users to hide data holders' inputs. The secret key of the two different third-parties is shared between data holders and the resulting cipher text is used as anonymized data for multiview point based clustering process. For the next encryption process, cipher text generated in the previous step is used as the message (plaintext) to be encrypted which yields the next random number as a result. The communication cost is analyzed between the Advanced Encryption Standard (AES) (İnan *et al*., 2007), Diffie Hellman Key Exchange Algorithm (DHKEA), Modified Diffie Hellman Key Exchange Algorithm (MDHKEA) and proposed RBFHE.

Figure 3 implies the communication cost of proposed RBFHE increases due to increasing amount of pair-wise entity comparisons with the existing methods such as MDHKEA, DHKEA and AES. Adult dataset containing 10 K entities is evenly distributed among data holders in these tests. It shows that the communication cost of the proposed RBFHE increases dramatically in our system due to secure comparison and communication cost of the existing methods is negligible compared to RBFHE.

Figure 4 implies the communication complexity of proposed RBFHE increases due to increasing amount of pair-wise entity comparisons with the existing methods such as MDHKEA, DHKEA and AES. House data set containing 10K entities is evenly distributed among data holders in these tests. It shows that the communication cost of the proposed RBFHE increases dramatically in our system due to secure comparison and communication cost of the existing methods is negligible compared to our proposed RBFHE for the same reason.

Figure 5 shows that the running time taken of clustering algorithm such as K-Means algorithm, Fuzzy C Means (FCM) clusters, Gaussian Firefly Algorithm (GFA) and MVBAT Clustering. The way of clustering the proposed MVBAT Clustering is different by means of similarity matrix and dissimilarity matrix, then two criteria function two criteria functions I_o&I_J is objective value for multiview point based clustering .The running time of the MVBAT is does not exceeds the normal time taken when compare to existing K-means algorithm, FCM, GFA for clustering the adult data set.

Figure 6 shows that the running time taken of clustering algorithm such as K-Means algorithm, FCM clusters, GFA and MVBAT Clustering. The way of clustering the proposed MVBAT Clustering is different by means of similarity matrix and dissimilarity matrix, then two criteria function two criteria functions I_o&I_J is objective value for multiview point based clustering .The running time of the MVBAT is does not exceeds the normal time taken when compare to existing K-means algorithm, FCM, GFA for clustering the house data set.

F-measure: The F-Measure quantifies how well a clustering that combines the precision and recall and constitutes a well-accepted and commonly used quality measure for automatically generated document clustering's. Let D represent the set of data matrix and let $C = \{C_1, \ldots, C_k\}$ be a clustering of D. Moreover, let $C^* = \{C_1^*, \ldots, C_l^*\}$ designate the reference partitioning. Then the recall of cluster j with respect to partition $i, rec(i,j)$, is defined as $|C_j \cap C_i^*|/|C_i^*|$. The precision of cluster j with respect to partition i, $prec(i,j)$, is

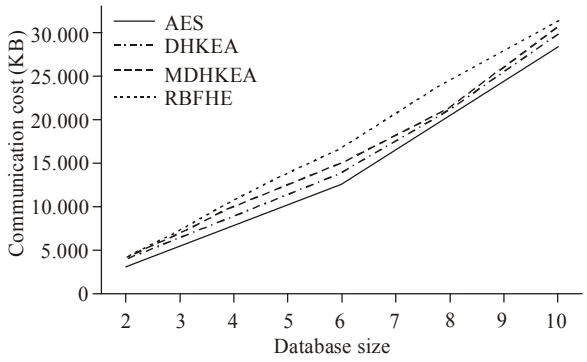

Fig. 3: Communication cost vs. methods for adult data set

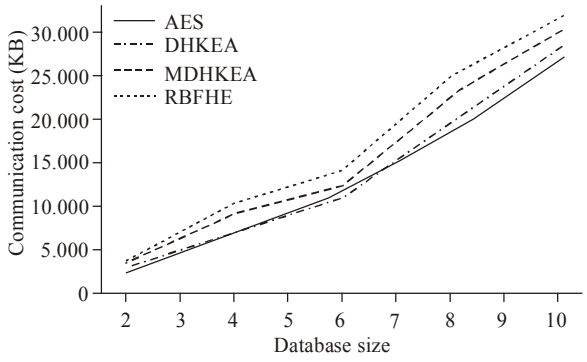

Fig. 4: Communication cost vs. methods for house data set

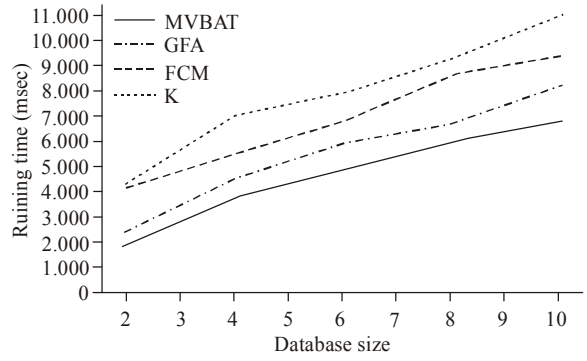

Fig. 5: Running time vs. clustering methods for adult data set

defined as $|C_j \cap C_i^*|/|C_j|$. The F-Measure combines both values as follows:

$$F_{i,j} = \frac{2}{\frac{1}{prec(i,j)}+\frac{1}{rec(i,j)}} \quad (32)$$

Based on this formula, the overall F-Measure of a clustering C is:

$$F = \sum_{i=1}^{l} \frac{|C_i^*|}{|D|} . \max_{j=1,\ldots k}\{F_{i,j}\} \quad (33)$$

Clustering results are evaluated using F-measure parameter and match point between the four raw cluster structures, results are demonstrated in Fig. 7 for adult dataset, it shows that the F measure accuracy of the

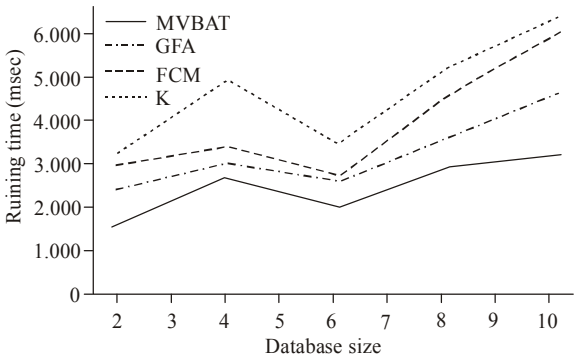

Fig. 6: Running time vs. clustering methods for house data set

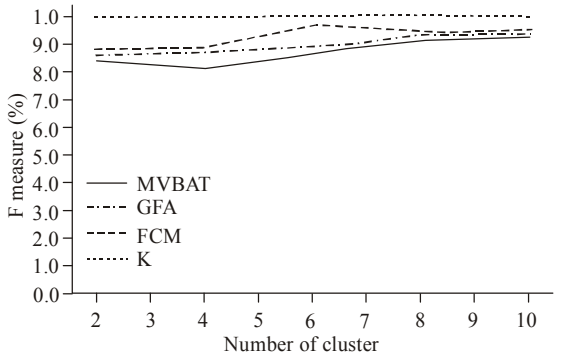

Fig. 7: F measure accuracy for clustering methods in adult data set

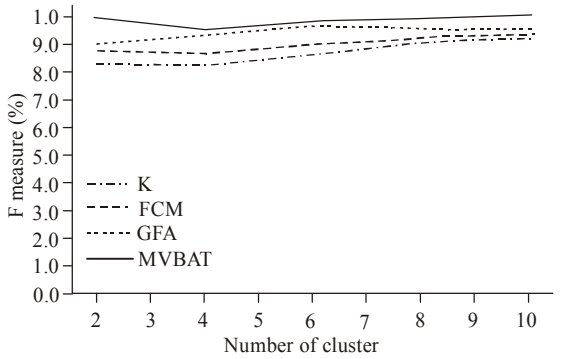

Fig. 8: F measure accuracy for clustering methods in house data set

proposed MVBAT clustering have higher value than the existing clustering methods. Since proposed work additionally multiview point based similarity measurement is performed when compare to existing methods.

In contrast to evaluate the clustering accuracy of four clustering methods separately measured using F-measure parameter and match point between the three raw cluster structures, results are demonstrated in Fig. 8 for house dataset, it shows that the F measure accuracy of the proposed MVBAT clustering have higher value than the existing GFA, FCM and K means clustering methods, proposed study additionally multiview point

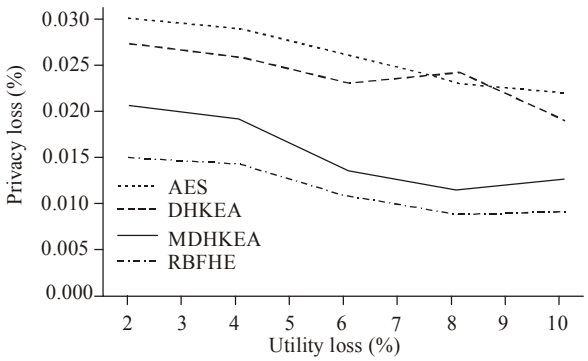

Fig. 9: Privacy loss vs. utility loss comparison for adult dataset

based similarity measurement is performed based on two criteria functions I_o& I_J.

Our results demonstrate the similarity between the privacy-utility loss in horizontal data partitioning data in adult dataset and house dataset for AES, DHKEA, MDHKEA methods and proposed RBFHE ,it shows that RBFHE methods provides substantially better data utility than existing encryption methods, the results are demonstrated in Fig. 8 and 9.

CONCLUSION AND RECOMMENDATIONS

In this study a novel privacy preserving multiview point based BAT clustering methods for privacy preserving clustering over horizontally partitioned data and the data disambiguation problem is solved by using RGSGK and anonymizes the data by using RBFHE encryption technique with secure key. RBFHE enclosed the data records of the data holders and key values are generated to perform anonymization process. MVBAT clustering two criteria functions I_o&I_J have been introduced to measure the similarity and dissimilarity values for data points for encrypted data samples from RBFHE. This two criteria functions I_o&I_J is considered as the objective function for clustering process in BAT algorithm. The quality of the resultant clusters from MVBAT can be easily measured and conveyed to data owners without any leakage of private information. Experimentation results of the proposed MVBAT is compared with other state-of-the-art clustering methods that use different types of similarity measure, on a UCI machine learning datasets such as adult dataset and house dataset and under different evaluation metrics, thus proposed MVBAT improves F-measure, less running time since MV similarity measurement is performed. Future research should examine the possibility of applying our method to vertically partitioning data clustering and perform same work under semi supervised clustering by considering unlabeled data matrix samples.

REFERENCES

Bach, F.R., 2008. Graph kernels between point clouds. Proceeding of the 25th International Conference on Machine Learning (ICML'08), pp: 25-32.

Brakerski, Z. and V. Vaikuntanathan, 2011. Fully homomorphic encryption from ring-LWE and security for key dependent messages. Proceeding of the 31st Annual Conference on Advances in Cryptology (CRYPTO'11), pp: 505-524.

Brakerski, Z., C. Gentry and V. Vaikuntanathan, 2012. Fully homomorphic encryption without bootstrapping. Proceeding of the 3rd Innovations in Theoretical Computer Science Conference (ITCS, 2012), pp: 309-325.

Bunn, P. and R. Ostrovsky, 2007. Secure two-party k-means clustering. Proceeding of the 14th ACM Conference on Computer and Communications Security, ACM, pp: 486-497.

Clifton, C., M. Kantarcioglu, J. Vaidya, X. Lin and M.Y. Zhu, 2003. Tools for privacy preserving distributed data mining. SIGKDD Explor., 4(2): 1-7.

Elisa, B., N.F. Igor and P.P. Loredana, 2005. A framework for evaluating privacy preserving data mining algorithms. Data Min. Knowl. Disc., 11:121-154.

Frank, A. and A. Asuncion, 2010. UCI Machine Learning Repository. School of Information and Computer Science, University of California, Irvine, CA. Retrieved from: http://archive.ics.uci.edu/ml.

İnan, A., S.V. Kaya, Y. Saygın, E. Savaş, A.A. Hintoğlu and A. Levi, 2007. Privacy preserving clustering on horizontally partitioned data. Data Knowl. Eng., 63(3): 646-666.

Jeckmans, A., Q. Tang and P. Hartel, 2012. Privacy-preserving collaborative filtering based on horizontally partitioned dataset. Proceeding of the IEEE International Conference on Collaboration Technologies and Systems (CTS, 2012), pp: 439-446.

Kantarcioglu, M. and C. Clifton, 2004. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE T. Knowl. Data En., 16(9): 1026-1037.

Lee, D. and J. Lee, 2010. Dynamic dissimilarity measure for support based clustering. IEEE T. Knowl. Data En., 22(6): 900-905.

Lindell, Y. and B. Pinkas, 2000. Privacy preserving data mining. Proceeding of Advances in Cryptology (Crypto'00). LNCS 1880, Springer-Verlag, pp: 20-24.

Liu, J., J.Z. Huang, J. Luo and L. Xiong, 2012. Privacy preserving distributed DBSCAN clustering. Proceedings of the Joint EDBT/ICDT Workshops, ACM, pp: 177-185.

Mangasarian, O.L., 2012. Privacy-preserving horizontally partitioned linear programs. Optim. Lett., 6(3): 431-436.

Ramon, J. and T. Gartner, 2003. Expressivity versus efficiency of graph kernels. Proceeding of the 1st International Workshop on Mining Graphs, Trees and Sequences, pp: 65-74.

Vaidya, J. and C. Clifton, 2003. Privacy-preserving k-means clustering over vertically partitioned data. Proceeding of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, D.C., pp: 206-215.

Verykios, V.S., E. Bertino, I. Nai Fovino, L. Parasiliti Provenza, Y. Saygin and Y. Theodoridis, 2004. State-of-the-art in privacy preserving data mining. SIGMOD Rec., 33(1): 50-57.

Xue, A., D. Jiang, S. Ju, W. Chen and H. Ma, 2009. Privacy-preserving hierarchical-k-means clustering on horizontally partitioned data. Int. J. Distrib. Sens. N., 5(1): 81.