IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

A Survey on Privacy Preserving Machine Learning Techniques for Distributed Data Mining

S. B. Javheri[1*], U. V. Kulkarni[2]

[1*]Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune, India
[2]Department of Computer Science and Engineering, S. G. G. S. I. E. & T., Nanded, India

[*]*Corresponding Author: sbjavheri@gmail.com, Tel.: 09763213131*

Available online at: www.ijcseonline.org

Accepted: 10/Jun/2018, Published: 30/Jun/2018

***Abstract*—** In the age of computer driven decision making the Data Science has become a vital important area for the parties storing the data. For the efficient use of available data; users need to excel in better Data Mining through robust Machine Learning techniques. Data mining applications are intensively used in government and corporate sector to analyze data for prediction, pattern recognitions, and classification. Accuracy of data mining algorithm depends on volume of training data. Advancement in computer and communication technologies allowed distributed computing environment, where multiple clients/parties can conduct joint learning process by incorporating distributed data. Distributed data may be arbitrary partitioned among parties. Recent data mining application uses power of cloud computing to execute complex computation involved in learning process. Despite of these advancements, individual or organizations holding data are reluctant to share their sensitive data due to fear of privacy breach and losses. Privacy Preserving Data Mining (PPDM) is solution to protect personal information while sharing it in distributed environment. Privacy preservation is achieved by data randomization or encryption techniques. Robust security to personal information and more accuracy in mining applications, offer more popularity to privacy preservation encryption techniques than data randomization techniques. Homomorphic encryption is one of the popular encryption techniques, where users can perform operations on cipher text; the results are similar to operations on their respective plaintexts. In present survey paper some data mining techniques like- ANN, RDT, SVM and Deep Learning, based on distributed partitioned data such are reviewed in special context to privacy preservation.

***Keywords*—** Data mining, Privacy preservation, homomorphic encryption, ANN, RDT, SVM, Deep Learning

I. INTRODUCTION

In present era, with the use of better Machine Learning Techniques data mining has become very exciting research domain in information technology and communi-cation industry. Data mining has many applications such as business intelligence, medical diagnostic systems, image processing, web search, scientific discoveries[1]. Machine learning techniques such as Artificial Neural Network (ANN), Random Decision Tree (RDT), Support Vector Machine (SVM) and Deep learning are widely used in data mining to explore information from distributed data, such as finding interested patterns, association in between entities, effect of specific entity on result, prediction, classification [1][2].

Traditional data mining applications uses data generated and maintained by single source for training the algorithms. Accuracy of data mining applications depends on volume of data samples used in training[3]. Advancement in computing allows training of data mining algorithms on distributed data samples, where two or more parties share their data for training and execute collaborative learning process. Performance and accuracy of distributed data mining applications are better than traditional data mining applications, due to training of learning algorithm on large volume of data samples shared by multiple parties[3][4].

Collaborative learning improves performance of financial agencies customer's credit evaluation system[5][4], centralize or distributed medical diagnosis system designed by sharing patients information to train diagnostic system on records of large number of patients[6][7]. A perfect buying pattern prediction system for malls is outcome if the system is trained on multiple records[6]. Sensor devices and Radio Frequency Identification (RFID) devices collect large volume of heterogeneous data in distributed environment[8][9]. Many data mining applications use services of these devices e.g. Smart city[9].
Cloud computing, an emerging technology, provides services for storage, computation, and infrastructure as per clients demand in distributed client-server environment. Cloud server provides additional resources as per the requirement of clients to perform their operations[10][11]. Cloud computing

executes collaborative learning process conveniently by executing complex operation and obtain result within time[3][9][12].

Despite, understanding advantages of distributed learning and advancement in computer networks, individuals and organizations are still reluctant to share their data in distributed environment[3][4]. Many individuals and organizations are concern about disclosing of their personal and proprietary information such as health record, economic information and research information[1][9]. In 2002, people from Japan and in 2003 people from US forced to stop government initiated record keeping and data mining program due to concern of personal information breach[13]. Sometimes data mining applications infer and disclose personal information of customers[1]. Financial agencies, corporate sector, organization found reluctant to share their customers' data to unknown data processing entities[4]. In some countries, Government Acts prohibit medical practitioners to share their patients' record to anyone without their permissions. Such as USA Govt act of Health Insurance Portability and Accountability(HIPAA) prohibits distribution of medical data of their patients[14]. Many users get awareness about the fact of misuse of their personal data through media. The most recent example is of Facebook. Such data breach is discovered after several months, even not by data owner, but some other users of that data[1].

Unavailability of a centralized trusted learner makes participants reluctant to share their data or execute collaborative learning process[15]. Participants either delete some information or provide false information due to privacy concern in providing private or proprietary information[6]. Now it is the responsibility of researchers to provide assured privacy preservation to users' data in collaborative learning process[1][3][6].

Privacy preserving data mining (PPDM) provides suitable solution where participants protects their personal information by encryption or modification and then shares in collaborative or centralized learning process[3][12]. Homomorphic encryption permit operations to be performed on ciphertext, which returns similar results if the operations are to be performed on their respective plain texts[16]. This property of homomorphic encryption made it popular in privacy preserving data mining, where direct operations on encrypted data are require[3][9].

The latter part of this paper includes Section 2: a discussion on data mining process and privacy preservation techniques. Section 3, discusses the data mining techniques such as ANN, RDT, SVM and Deep computation model with privacy preservation. Section 4 gives overall conclusion of Privacy Preserving Machine Learning Techniques.

II. DATA MINING AND PRIVACY PRESERVATION

A) DATA MINING PROCESS

Data mining process also recognized as "Knowledge Discovery from Data (KDD)" [2], normally executed in four steps, in each step either a person or an organization participates in typical data mining process. These users care about the security of sensitive information [1].

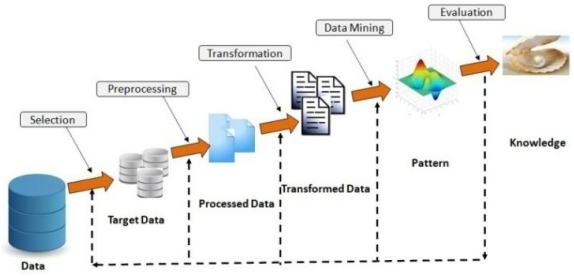

Figure1: KDD Process

- **Data Preprocessing –** Data provider provides expected data for data preprocessing in first step of KDD.
- **Data Transformation –** Data collector selects data features and transforms data in second step.
- **Data Mining –** Data miner applies intelligent machine learning techniques to identify interested data pattern in third step.
- **Pattern evaluation and presentation -** Decision maker represent pattern in final step of KDD.

Once the data is handed over to other party then there is no control on how the receiver uses that data. Therefore it is a need of time to design privacy preserving data mining model will preserve the privacy of participants' data in all mining steps [1].

B) Data Distribution

Recent advancements in internet technology makes possible the availability of centralized or distributed data among participants. Therefore, privacy preserving data mining process is required in centralized or distributed data processing [1].

1) Centralized Database: Participants outsource required data using common database schema to central server in building centralized learning model.

Dataset DB					
a_1	a_2	a_3	a_4	a_5	a_6
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}

Figure 2: Centralized Database

2) Distributed database: distribution of data among multiple parties, in horizontally, vertically [1] and arbitrarily[5].

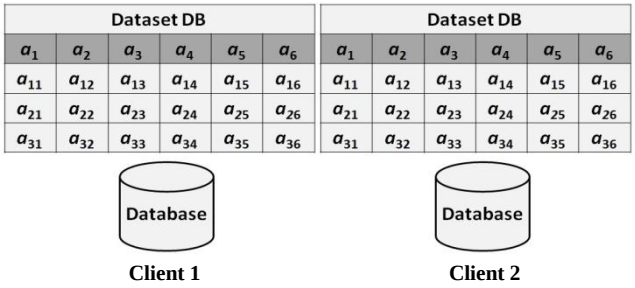

Figure 3: Distributed Database

- **Horizontal partition:** Participants hold data values for some rows of complete database.
- **Vertically partition:** Participants hold data values for some columns of database.
- **Arbitrarily partition:** Participants hold data values for arbitrary way.

C) Privacy Preserving Techniques

Privacy preservation techniques protect user's personal information by converting plain values into cipher text. Privacy preserving data mining application introduced in year 2000 with two different techniques- data randomization and cryptography.

1. Randomization-based Approaches –

Agarwal[16] invented and used randomization (data perturbation) techniques for privacy preservation by adding noise to the source data.

2. Cryptographic-based Approaches –

Secure Multi-party Computation (SMC), a seminal work by Yao [17] and Goldreich [18] are pioneers to numerous privacy preserving data mining techniques. Lindell and Pinkas [19] have applied cryptographic tools to build secure decision tree classifier, which is proved efficient than data perturbing method. Privacy preservation and accuracy of cryptographic techniques is good, but they are difficult to handle with large data sets because due to demanding of more resources.

With the directions of these two land mark contributions, number of privacy preservation data mining systems are further designed by researchers in supervised and unsupervised learning process.

After the invention of well-known RSA encryption, Rivest, Adleman and Dertouzous have first time demonstrated homomorphic encryption property of RSA [21][22]. Homomorphic encryption results of operations performed on ciphertexts are equal to operation on their respective plaintexts without decryption of ciphertexts [15], this property makes it more useful in privacy preserving data mining. Homomorphic encryption, a cryptographic-based technique, has better ability to execute secure multiparty computation [20], Homomorphic encryption is classified into –

a) Additive Homomorphic Encryption –

Product of two ciphertexts Ek(PT1) and Ek(PT2) is equal to addition of their plaintext.

Ek(PT1 +PT2) = Ek (PT1) * Ek (PT2)

E.g. Okamoto-Uchiyama [23], Pallier [24]

b) Multiplicative Homomorphic Encryption -

Multiplication of two ciphertexts Ek(PT1) and Ek(PT2) equal to product of their plaintexts.

Ek(PT1 *PT2) = Ek (PT1) * Ek (PT2)

E.g. Rivest et al.[21] and ElGamal [25]

III. Privacy Preservation Machine Learning Techniques

Data classification is important data mining applications where users examine the data and extract its categories as classes. A classifier gets developed in two steps- training phase and testing phase. A classifier is designed with machine learning technique such as - ANN, RDT, SVM, Naïve Bayesian Classification[2] and Deep Learning Model [8] etc.

A) Privacy Preserving Back-Propagation Neural Network

Neural Networks are usually used for regression and classification. Back propagation neural network (BPN) is effective learning method.

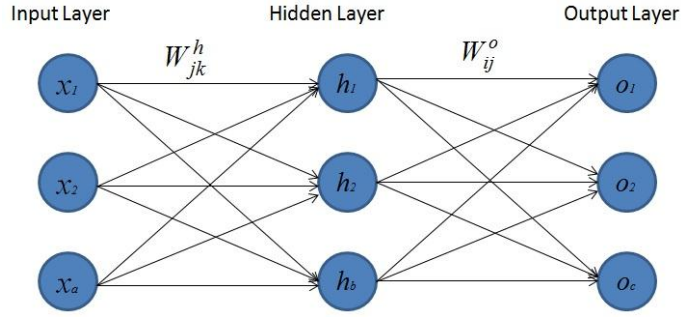

Figure 4:Three layer a-b-c Back-Propagation Neural Network

Back propagation neural network with three layer shown in figure 4. Vector { $x_1, x_2, \ldots, x_a$ } vector { $h_1, h_2, \ldots, h_b$ } and vector {$o_1, o_2, \ldots, o_a$} to represent the values of input layer nodes, hidden layer nodes and output layer nodes, respectively.W^h_{jk} is used as weight connecting to the input layer node k and the hidden layer node as j. W^o_{ij} indicate the weight linking as j and the output layer node as i, where $1 \leq k \leq a$, $1 \leq j \leq b$, $1 \leq i \leq c$.

BPN network learning process operates in feed forward and back-propagation stage.

- **Feed Forward Stage:** Applying the values at previous layer, weights and sigmoid function feed forward stage calculates respective hidden and output layer neuron values.
- **Back-Propagation Stage:** In this stage algorithm checks the error by calculating difference between actual output values and target values which is required below threshold. Otherwise, accordingly all the weights will be modified by equations –

$$\Delta w_{ij}^{o} = -(t_i - o_i)h_j$$

$$\Delta w_{jk}^{h} = -h_j(1-h_j)x_k \sum_{i=1}^{c}[(t_i - o_i) * w_{ij}^{o}]$$

To get better accuracy of learning results, multiple parties may work together on the union of their data sets through combined error back-propagation neural network learning [3][5]. We need a feasible solution wherein participants who lack mutual trust can participate in internet-wide collaborative learning without exposing their respective personal data.

Schlitter[28] designed a privacy protecting BPN network learning on horizontally partitioned data with two or more parties. Scheme does not protect intermediary results, which may contain sensitive information. Chen and Zhong[14] developed BPN network learning algorithm with privacy preserving for vertically separated data with two party scenario. Bansal et al[5] proposed arbitrary partitioned data for two party scenarios. Jiawei Yuan[3] propose a scalable collaborative learning model which maintains privacy preservation with learning on arbitrary partitioned data.

Scheme Overview

Privacy preserving classification system designed by Jiawei Yuan[3] has three components- trusted authority (TA), parties involved in collaborative learning and server deployed in cloud. TA generates encryption/decryption keys using Boneh–Goh–Nissim (BGN) homomomorphic encryption[29] and does not participate in any actual computation. Participating parties owns arbitrary partitioned data, and cloud server perform computation involved in multiparty BPN network.

Data Partition

Back-propagation neural network using 3-layer (a-b-c configuration) is used, and it is extendable to multilayer neural network. *N* samples learning dataset denoted as vector $\{x_1^m, x_2^m, \ldots, x_a^m\}$, $1 \le m \le N$, is randomly partitioned amongst Z parties($Z \geq 2$). Each P_s party, $1 \le s \le Z$, holds x_{1s}^m, $x_{2s}^m, \ldots, x_{as}^m$

$$x_{11}^m + x_{12}^m + \cdots + x_{1Z}^m = x_1^m$$

$$x_{a1}^m + x_{a2}^m + \cdots + x_{aZ}^m = x_a^m$$

One party holds some attributes in a sample, which are not available with other parties and vise a versa $\{x_1^m, x_2^m, \ldots, x_a^m\}$, $1 \le m \le N$, – if P_s possessed x_k^m $1 \le k \le a$, then $x_{ks}^m = x_k^m$ otherwise $x_{ks}^m = 0$

Privacy Concern

BGN homomorphic encryption executes one multiplication and *n* number of addition over ciphertexts. Jiawei Yuan has modified BGN scheme, so that it will be able to decrypt large numbers by considering message as bit string. For secure computation of consecutive multiplication a secret sharing scheme is used with securely decrypt intermediate products or scalar product in BPN network learning.

Neural Network Learning Process

For collaborative learning, all parties should together perform operations defined in the feed forward and back-propagation stage. During each learning step only final learned network revealed to participant.

At the beginning, all parties initialize random weights W_{jks}^h and W_{ijs}^o to each P_s party and commonly agree on - max numbers of learning iteration, learning rate η, target values, error threshold. Trusted Authority (TA) generates system master key q and distributes its random share to P_s party for encryption. Participating parties encrypt all values using this random share key. In feed forward phase, as per accuracy requirements of each party they agree on approximation for the sigmoid function. Computation of sigmoid function over ciphertext is highly impractical therefore; Maclaurian series expansion to approximate the sigmoid function which is used as -

$$\frac{1}{1+e^{-x}} = \frac{1}{2} + \frac{x}{4} - \frac{x^3}{48} + \frac{x^5}{480} + O(x^6)$$

Obtain random shares h_{js} and o_{is} for value of hidden and output layer node respectively. After the feed forward stage, all the parties work jointly to verify whether the network has achieved the error threshold or not. If not then, they proceed for back-propagation stage to adjust weights. For the weights of each output layer node W_{ij}^o, each *Ps* obtains random shares of the changes in weights denoted as ΔW_{ijs}^o for ΔW_{ij}^o by using secure scalar product and sum algorithm. To compute the changes in the weight ΔW_{ij}^o of each hidden layer node proposed scheme calculations are –

$\sum_{i=1}^{C}\left[(t_i - o_i) * w_{ij}^o\right] x_k \sum_{i=1}^{C}\left[(t_i - o_i) * w_{ij}^o\right]$, $-h_j(1-h_j)$ and W_{jk}^h

Each P_s obtains the random shares $(\mu_s, ks, vs, \Delta w_{jks}^h)$ of each item using Algorithms 1) Secure Scalar Product and Addition, 2) Protected Share of Scalar Product and Sum. Finally, P_s revises its weights with learning rate $\eta \Delta W_{jk}^h$ and its shares.

Jiawei Yuan et.al[3] design scalable multiparty collaborative BPN network learning over randomly separated data. This scheme guarantees about privacy and efficiency with lower

computation/communication costs than existing one. It supports multiparty secure scalar product and decryption of large messages in BGN cryptosystem.

B) Privacy Preservation Random Decision Tree Approach

Random Decision Tree (RDT) is inductive learning based approach to finding most favorable hypothesis. First it builds *N* random decision trees, and then updates each node with scanning and training samples. RDTs are used in many data mining applications such as classification, regression, ranking and multiple classifications. RDT is successful non-parametric density estimation and can be explained via high order statistics such as moments. Use of multiple RDTs is effective in distributed or parallel tasks.

For data classification RDT operates in training phase and classification phase. Training phase builds decision trees by selecting any feature randomly without data values, and stop building tree when depth of tree is equivalent to half of the features of data set. Then apply training data to revise information of every node. In classification phase, only leaf nodes declare classification result of unknown data samples. RDT's property of random structure is used as data perturbation technique for privacy preservation, where only leaf nodes need encrypted by cryptographic based approach [31].

Construction of RDT

Consider case of two party distributed weather dataset, where data is partitioned horizontally or vertically.

If data is horizontally partitioned and distributed among *k* parties (P_1, P_2,... P_k) then each party knows database schema and class labels, but where data is vertically partitioned then, a party who posses attribute knows sections of database schema and class labels.

As per RDT building process two RDTs are constructed as –

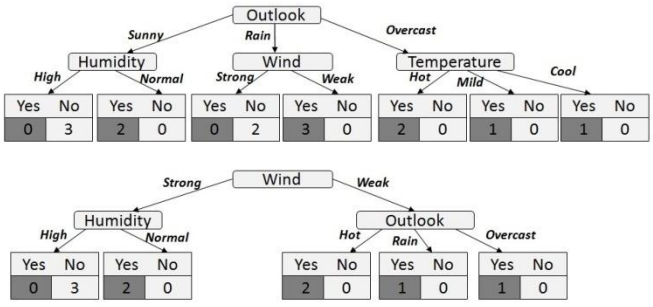

Figure 5. Random decision trees for weather data set

A new instance (sunny, mild, normal, weak) is predicated (2,0) by first RDT and (1,0) by second RDT respectively, results in overall class distribution is (1.5,1).

Wang et al. proposed solution for vertically partitioned data based on passing of transaction identifier. This is based on transaction execution path wherein the parties can guess about the attributes hold by other parties. Du and Zhan proposed a two party privacy preserving decision tree solution for vertically partitioned data. Vaidya proposed a multi-party privacy preserving decision tree solution for vertically partitioned data.

Horizontally partitioned data-

When data is horizontally partitioned then all parties independently construct RDTs, by cooperatively and securely share leaf node values.

	party p_1		party p_2		
	outlook	**temperature**	**humidity**	**windy**	**play**
party p_1	sunny	hot	high	weak	no
	sunny	hot	high	strong	no
	overcast	hot	high	weak	yes
	rainy	mild	high	weak	yes
	rainy	cool	normal	weak	yes
	rainy	cool	normal	strong	no
	overcast	cool	normal	strong	yes
party p_2	sunny	mild	high	weak	no
	sunny	cool	normal	weak	yes
	rainy	mild	normal	weak	yes
	sunny	mild	normal	strong	yes
	overcast	mild	high	strong	yes
	overcast	hot	normal	weak	yes
	rainy	mild	high	strong	no

Table 1. Distributed data set for Weather

In horizontal partitioned data set authors consider two cases: 1) Each party knows complete tree structure, and 2) some parties don't know the complete structure. In second case parties don't calculate leaf node values, so scenario is quite complicated.

When every party knows tree structure, again there are three situations based on global class vector for each leaf node – 1a) all parties know global class vector, 1b) only party, who passes the tree, know global class vector 1c) unknown to all parties.

Case 1a

For privacy preservation, if any one party find out that common tree structure under construction reveals too much information and about to identity of specific instance class, it will reject that tree structure and start new process for building tree. There is a flaw in this method, if any party satisfied with ongoing tree structure, then it can guess about class data instances possessed by unsatisfied parties. After agreement on final tree structure all parties first calculate leaf node values at local level and using secure share protocol compute global values[30].

Case 1b

Parties, which hold the tree, execute secure sum protocol to calculate values for leaf node. Party P_i, not possessing the tree, want to classify new instance first generate public-private key pair using additive homomorphism [zurich] with threshold cryptography. P_i Sends public key to all tree owner parties. Tree owner parties encrypt leaf nodes and send encrypted tree to P_i, multiply all these encrypted vectors and got classification result through threshold cryptography.

Case 1c

Most difficult case, where in computation of secure sum, either parties having share are participate or parties which own encrypted tree participates. First parties agree on tree structure based on secure electronic voting. Each party locally calculates leaf node values for class distribution. Encrypt it using additive homomorphism and threshold cryptography and send for global computation. To decrypt these values other participant parties require. Each party build global tree in decrypted form.

To obtain the encrypted sum of the class distribution vectors, in classification of a new instance, party first identifies all leaf nodes instance received and multiplies encrypted class vector sum components together. This result jointly decrypted and averaged to get the actual class distribution.

Vertically Partitioned Data

In vertically separated data every party holds some set of attributes related with similar entity. Two cases are considered either parties share information of attributes to build trees independently or nothing share in between parties and independent trees are constructed in distributed form at party level.

Consider the case for the parties build *m* independent trees with level *n*/2 where *n* is total number of attributes in overall schema R, using secure sum protocol. After tree structure is ready node values will be updated by scanning training data sample in distributed and privacy preserving manner. Tree construction process start by selecting any random site *j* with any one attribute posses by *j*. if attribute is numeric then a split point π is decided to create left and right child of tree. For categorical attribute recursive process called for each value. Additive homomorphism is used to update class value in encrypted for at leaf node.

To update class value first site which own attribute constructed a class vector of attributes with random encryption for '1' for its attribute for rest random encryption for '0' using Paillier encryption[24]. This will helpful to privacy preservation of which class value is updating. Then each distributed random tree is traversed with encrypted class values by passing control from site to site. At the end specific leaf node value is updated.

Classification of new instance predicted by a distributed algorithm by computing aggregate probability outputs of multiple RDTs. Classification process transfer new instance from one RDT to another based on its attributes information.

A multi-party privacy preserving solution is developed for horizontally and vertically partitioned data. Randomness in tree structure provided strong privacy and less computation cost. Data size not affect on classification performance of RDTs.

C) Privacy Preservation Multi-Class Support Vector Machines (SVMs)

Support Vector Machine (SVM) having good mathematical base and better generalization of data. Therefore SVM is extensively used machine learning for data classification, text mining, computer vision, natural language processing, bioinformatics and many applications. There is training phase and testing phase in SVM classification. Training phase operates on, known input data samples provide as input so classification attributes are retrieved. In testing phase classifier accept unknown data samples and labelled with specific class. SVM introduce to classification of two-class problem, this is extended for multi-class problem by dividing multi-class problem in multiple two-class problem.

With a training set of points $\tilde{x}_i \subset \mathbb{R}^n, i = 1, \dots, N$, where every point $\tilde{x}_i$ fit in to any one of class from two classes denoted by a label $\tilde{x}_i \subset \{-1, +1\}, i = 1, \dots, N$. Two class problems further break up into linear classification and non-linear classification.

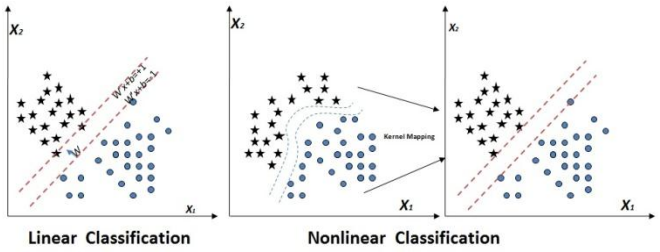

Figure 6 Support vector machine classifications

In Linear classification training phase data samples linearly classified into two parallel hyperplane **w′x** +b= -1 and **w′x** +b= +1, where **w** and *b* are the parameters for classification obtained in training process. In testing phase, unknown test sample $\tilde{t} \subset \mathbb{R}^n$ after normalization, put into function for classification

$$f(t) = sign(w't + b) = sign(\sum_{s \subset S} \alpha_s y_s x'_s t + b)$$

Where $f(t) \in \{-1, +1\}, \alpha_i \ i = 1, \dots, N$ are langrangian variables and x_s s=1,…,|S|are support vectors. If $f(t)$=+1 then the test sample $\tilde{t}$ belongs either *+ve* class or belongs to *–ve* class. Decision function $d(t)$ can be derived as -

$$d(t) = w't + b = \sum_{s \subset S} \alpha_s y_s x'_s t + b$$

Where, $\boldsymbol{w'x}+b= 0$ denotes the decision-hyperplane which lies between the two hyperplanes(i.e. $\boldsymbol{w'x}+b= -1$ and $\boldsymbol{w'x}+b= +1$)

A non-linear classification process is similar to linear classification, where dot product (i.e. $x'_s t$) is replaced by a various non linear kernel functions, which maps data samples into higher dimensional feature space.

With consideration of polynomial kernel, dot product between $\mathbf{x}_s$ and $\mathbf{t}$ can be replaced as –

$$x_s t \Rightarrow K(x_s t, 1)^p = (x_s t + 1)^p$$

Multi-Class Problem:

A multi-class problem can be decoupled into multiple two-class problems. There are two different approaches for this one-versus-all (IVA) and one-versus-one(IVI).

One-versus-All Apprach-

Nc number of classes, framed into N_c number of sub-problems. Hence, we train *Nc* number of SVMs. For j^{th} sub-problem, SVM trained as a +ve class and remaining samples as –ve class. For a given normalized test sample, *t*, the matching class, *Mc* can be obtained in the testing phase, by modifying as-

$$M_c = \underset{j=1,\ldots,N_c}{\arg max} \left(\sum_{s \subset S} \alpha_s y_s (x'_s t + 1)^p + b \right)$$

One-versus-One Approach

for this approach, we train $\frac{N_c(N_c-1)}{2}$ number of subproblems, where SVM for each subproblem trained using data only from said two classes. If we consider training samples from two classes *i* (i. e. +ve class) and *j* (i.e. –ve class) for a given subproblem, then classification function can be written as –

$$f_{i,j}(t) = sign\left(\sum_{s \subset S} \alpha_s y_s (x'_s t + 1)^p + b \right)$$

Where subscript *i*, *j* denotes the variables associated with the $(i,j)^{th}$ subproblem. *t* can be obtained for the matching set of class for the given test sample by using majority voting approach as –

$$M_c = \underset{j=1,\ldots,N_c}{\arg max} \left(\sum_{j=1, j \neq i}^{N} f_{i,j}(t) \right)$$

Classification of Encrypted Data

In client-server classification model, server maintains training data samples and acquires classification parameters. Paillier cryptosystem is used for encryption of data values. Encryption of negative values performed under cyclic property of modulo arithmetic.

For two class classification client send encrypted values to server. Paillier system operate on integer values only, therefore a decision function is used to change values to nearest integer. This is helpful to maintain classification accuracy.

On server side first encrypted data samples normalized with parameters mean and deviation and then scaled. Server compute polynomial kernel in encrypted domain. If degree of polynomial p=1 then server not interacting with client for computation of kernel, whereas if $p \geq 1$ then using secure two-party computation server interact with client and compute kernel. To maintain privacy of server side parameters server sends mask kernel parameters with some mask and send to client. Client decrypt this values with its private key and raise decrypted values with degree *p* encrypt it, and send to server. Now server computes decision function using Paillier additive function i.e. product of all required parameters in encrypted form. To obtain the class sign from this result server calculate a variable *z*, if most significant digit of z=1 then test sample fit in to $class_+$ and if z=0 then *class_*.

A multi-class classification obtained by deviding multi-class problem into many two-class subproblem. In 1VA scenario server needs to find out decision parameter for each two-class subproblem in encrypted domain. Identify a subproblem with largest decision function value with test sample related with that class. Use recursive algorithm on *N* classes with secure two party computations to find out largest decision value of encrypted test sample.

In 1*V*1 approach each subproblem SVM trained using test samples of two classes. Based on voting protocol identify a class, which received maximum +ve guesses.

A client-server based privacy preserving classifier designed using Support Vector Machine (SVM) for two-class and Multi-Class classification. Paillier cryptosystem is used for additive homomorphic encryption; a secure protocol is used to obtain sign of encrypted value. Classification accuracy of encrypted data and plain data is same. In Semi honest model, only clients may become adversary to identify SVM class definition vector.

D) Privacy Preserving Deep Computational Model

Tensor auto-encoder (TAE) is used to model non-linear distribution of heterogeneous big data[8]. Deep computation model used stacked architecture of several tensor auto-encoders.

Suppose two tensors, $X \in R^{I_1 * I_2 * \ldots I_N}$ and $H \in R^{J_1 * J_2 * \ldots J_N}$ represents values for input layer and hidden layer nodes respectively. TAE encoder function relates input values to hidden layer values as – $H = f_\theta(W^{(1)} O X + b^{(1)}$. TAE represent as $Z^{(2)}_{j_1, j_2, \ldots, j_n} (1 \leq j_i \leq J, 1 \leq i \leq n)$ and for output layer input values represent as $Z^{(3)}_{i_1, i_2, \ldots, i_n} (1 \leq i_j \leq I_j, 1 \leq j \leq$

n). Activation value for input layer are decoder function relates hidden values to reconstruction Y as $Y = hw, b(X) = g_\theta(W^{(2)}OH + b^{(2)})$, where $\theta = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$ is parameter set, enccoder and decoder function use Sigmoid function $s_f(x) = 1/(1+e^{-x})$, $\odot$ denote the multi-dot product of two tensors. To retrieve maximum distribution from data tensor distance is used in reconstruction error. In reconstruction error, first terms represents average sum-of-squares error term and the second term is a regularization term or weight decay term. Input values of hidden layer $a^{(2)}_{j_1,j_2,\ldots,j_n}(1 \le j_i \le J_i, 1 \le i \le n$ and for hidden layer are) $a^{(3)}_{i_1,i_2,\ldots in}(1 \le i_j \le I_j, 1 \le j \le n)$.

Privacy preserving High-Order Back-Propagation

To train parameters of TAE Cloud server execute high-order back propagation algorithm in feed forward stage and back-propagation stage.

In feed forward stage to calculate values for hidden layer and output layer algorithm used training parameters, sigmoid function and previous layer values.

In back-propagation stage update weights using updating reconstruction error. For privacy preservation BGV encryption is used to encrypt data. BGV method is a LWE/RLWE based fully homomorphism encryption. Client first encrypt input data $\{x_1, x_2, \ldots x_a\}$, output data $\{t_1, t_2, \ldots t_c\}$ and initialized parameters $\theta = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$ as message and send this to cloud server for first round of iteration. Deep computation model increase its circuit size on each iteration, this will reduce efficiency of deep computation model. To overcome this issue, after each iteration server send results back to client for updating the parameters. Client decrypt results received from server and update parameter. Again client encrypt updated parameters and send for iteration. This step repeated until max iterations not crossed.

To reduce complexity of computation of encrypted data, exponentiation operation of sigmoid function replace with Taylor theorem of approximation. BGV property of Secure addition and Secure multiplication is used for addition and multiplication of encrypted data.

Efficient Privacy protecting deep computational model for big data feature learning on cloud computing is designed. Performance of the scheme will be improved by adding extra cloud servers. Presently performance is slightly less than non-privacy preserving deep computational scheme.

Table2. Summarized machine learning techniques used for privacy preservation classification.

	Machine Learning algorithm	**Data Distribution**	**Privacy Concerns**	**Method Description**	**Performance Measurement**	**Data mining model deployment**
J. Yuan and S. Yu[3]	Back-Propagation Neural Network	Distributed (Arbitrary Partitioned)	Semi honest model Complete Privacy preservation of input data and intermediate results.	BGN cryptosystem, SMC based Secure Scalar Product and Addition, Protected Share of Scalar Product and Sum	Learning time, communication cost, error rate comparison	Each party prepare common trained BPN network
Jaideep Vaidya et al.[31]	Random Decision Tree	Distributed (Horizontally and Vertically Partitioned)	Semi honest model, Semantically secure homomorphism and Threshold cryptography maintain data privacy, At classification time leakage of average sum from class distribution vector	Randomness in structure, Paillier cyptosystem, Secure Share Protocol Secure Sum Protocol	Build Tree, Classification Time Classification Accuracy	Independent RDTs constructed on each site for classification.
Y Rahulma Dhavan [11]	Support Vector Machine	Distributed	Client-Server interaction makes clients act as adversary	Semi honest Model, Paillier Cyrptosystem, Secure Two Party Computation	Classification accuracy, Computation time	Centralized Server build two-class and multi-class SVM classifier
Qingchen Zhang[8]	Deep Computatio n Model	Distributed	Only client can decrypt the values.	BGV fully homomorphic encryption, secure addition, secure multiplication	Classification accuracy,	Centralized server maintain deep computational model

IV. CONCLUSION

Privacy preserving techniques are successfully used with machine learning algorithms. This will boost collaborative learning process and people will be ready to share their data to design data mining applications with better efficiency, accuracy and less in computation and communication cost. Data mining applications on arbitrary partitioned data is a challenging task to many machine learning algorithms. Back-propagation is a versatile learning algorithms used independently and in emerging deep computational model. J Yuan[3] made progress on privacy preservation classification on arbitrary partitioned data using back-propagation neural network.

To obtain privacy preservation in machine learning algorithms require time consuming encryption and decryption operations on plain text. This results in slow operating speed of machine learning algorithms and also decreases in accuracy. There is further scope to deploy efforts required to reduce encryption operation overheads and increase accuracy.

Also in all above research work discussed, a major finding is that either in horizontal, vertical or arbitrary partitioned data researchers considers condition that if one party hold set of attributes from sample; there will be possibly no other parties holding same set of attributes for that sample. In distributed environment with arbitrary partitioned data, overlapping of attributes also required to be considered. There is further study required on overlapped attributes cases in arbitrary partitioned data.

References

[1] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1151–1178, 2014.

[2] J. P. J. Han, M. Kamber, *Data Mining: Concepts and Techniques*. San Mateo: Morgan Kaufmann, 2006.

[3] J. Yuan and S. Yu, "Privacy Preserving Back-Propagation Neural Network Learning Made Practical with Cloud Computing," *Tpds*, vol. 25, no. 1, pp. 212–221, 2014.

[4] Y. Zhang and S. Zhong, "A privacy-preserving algorithm for distributed training of neural network ensembles," *Neural Comput. Appl.*, vol. 22, no. S1, pp. 269–282, 2012.

[5] A. Bansal, T. Chen, and S. Zhong, "Privacy preserving Back-propagation neural network learning over arbitrarily partitioned data," *Neural Comput. Appl.*, vol. 20, no. 1, pp. 143–150, 2011.

[6] M. Chicurel, "Databasing the brain.," *Nature*, vol. 406, no. 6798, pp. 822–5, Aug. 2000.

[7] K. Flouri, B. Beferull-Lozano, and P. Tsakalides, "Training a SVM-based classifier in distributed sensor networks," *Eur. Signal Process. Conf.*, pp. 1–5, 2006.

[8] Q. Zhang, L. T. Yang, and Z. Chen, "Privacy Preserving Deep Computational Model on Cloud for Big Data Feature Learning," vol. 65, no. 5, pp. 1–11, 2015.

[9] R. L. Grossman, "The case for cloud computing," *IT Prof.*, vol. 11, no. 2, pp. 23–27, 2009.

[10] R. L. Grossman and Y. Gu, "Data Mining Using High Performance Data Clouds: Experimental Studies Using Sector and Sphere," *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 920–927, 2008.

[11] M. Rajarajan, K. Cumanan, S. Veluru, R. C.-W. Phan, and Y. Rahulamathavan, "Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud," *IEEE Trans. Dependable Secur. Comput.*, vol. 11, no. 5, pp. 467–479, 2013.

[12] J. Vaidya and C. Clifton, "Privacy-Preserving Data Mining: Why, How, and When," *IEEE Secur. {&} Priv.*, vol. 2, no. 6, pp. 19–27, 2004.

[13] "Your Rights Under HIPAA | HHS.gov." [Online]. Available: http://www.hhs.gov/hipaa/for-individuals/ guidance-materials-for-consumers/index.html. [Accessed: 13-May-2016].

[14] T. Chen and S. Zhong, "Privacy-preserving backpropagation neural network learning.," *IEEE Trans. Neural Netw.*, vol. 20, no. 10, pp. 1554–64, 2009.

[15] C. Gentry, "Fully Homomorphic Encryption without Bootstrapping," *Security*, vol. 111, no. 111, pp. 309–325, 2011.

[16] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *Proc. 2000 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '00*, vol. 29, no. 2, pp. 439–450, 2000.

[17] A. C. Yao, "Protocols for secure computations," *23rd Annu. Symp. Found. Comput. Sci. (sfcs 1982)*, pp. 1–5, 1982.

[18] O. Goldreich, S. Micali, and A. Wigderson, "How to Play any Mental Game," *Stoc '87*, pp. 218–229, 1987.

[19] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," *J. Cryptol.*, vol. 15, no. 3, pp. 177–206, 2002.

[20] J. Sen, "Homomorphic Encryption: Theory & Applications," *arXiv Prepr. arXiv1305.5886*, pp. 1–32, 2013.

[21] R. L. Rivest, L. Adleman, and M. L. Dertouzos, "On Data Banks and Privacy Homomorphisms," *Found. Secur. Comput. Acad. Press*, pp. 169–179, 1978.

[22] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Commun. ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[23] T. Okamoto and S. Uchiyama, "A new public-key cryptosystem as secure as factoring," *Advance Cryptology—EUROCRYPT' 1998, Lect. Notes Comput. Sci.*, vol. 1403, pp. 308–318, 1998.

[24] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes," *Adv. Cryptol. — EUROCRYPT '99*, vol. 1592, pp. 223–238, 1999.

[25] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *Inf. Theory, IEEE Trans.*, vol. 31, no. 4, pp. 469–472, 1985.

[26] C. Gentry, "Fully homomorphic encryption using ideal lattices," *Proc. 41st Annu. ACM Symp. Symp. theory Comput. STOC 09*, vol. 19, no. September, p. 169, 2009.

[27] Z. Brakerski, "Fully homomorphic encryption without modulus switching from classical GapSVP," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7417 LNCS, pp. 868–886, 2012.

[28] N. Schlitter, "A Protocol for Privacy Preserving Neural Network Learning on Horizontally Partitioned Data," 2008.

[29] D. Boneh, E. Goh, and K. Nissim, "Evaluating 2-DNF formulas on ciphertexts," *Theory Cryptogr.*, pp. 325–341, 2005.

[30] B. Schneier, "Applied Cryptography," *Electr. Eng.*, vol. 1, no. [32, pp. 429–455, 1996.

[31] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, "A Random Decision Tree Framework for Privacy-Preserving Data Mining," *IEEE Trans. Dependable Secur. Comput.*, vol. 11, no. 5, pp. 399–411, 2014.

Authors Profile

S. B. Javheri in pursed Bachelor of Engineering from North Maharashtra University, Jalgaon, Marashatra, India in 1998 and Master of Engineering from B. V. D. University, Pune, Maharashtra, India in year 2009. He is currently pursuing Ph.D. from S.G.G.S.I.E.& T., Nanded, Maharashtra, India and currently working as Associate Professor in Department of Computer Engineering , JSPM's Rajarshi Shahu College of Engineering, Pune since 2004. He has published 12 research papers in reputed international journals/conferences. His main research work focuses on information security, machine learning, Neural network. He has 19 years of teaching experience.

U. V. Kulkarni has obtained Bachelor of Engineering degree in Electronics from Marathwada University, Aurangabad, Maharashtra, India in 1987. He completed Master of Engineering in system software from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India in 1992. He has completed his Ph.D. in Electronics and Computer Science Engineering in 2002 from Swami Ramanand Teerth Marathwada University Nanded, Maharashtra, India. He is currently working as Professor and Head in Computer Science and Engineering Department at SGGSIE&T (Autonomous), Nanded, Maharashtra, India. He has received National Level Gold Medal and Computer Engineering Division Prize for the paper published in the Journal of Institution of Engineers, titled as Fuzzy Hypersphere Neural Network Classifier, May 2004 and the best paper award for the research paper presented in international conference held at Imperial College London, U.K., 2014. He has published many research papers in reputed National and International Journals. His areas of interest include Microprocessors, Data Structures, Distributed Systems, Fuzzy Neural Networks, and Pattern Classification. He has more than 30 years of teaching experience.

