A Data Mining Perspective in Privacy Preserving Data Mining Systems

Article · January 2014

CITATIONS
3

READS
310

4 authors:

Kumaraswamy S
KNS Institute Of Technology
9 PUBLICATIONS **9** CITATIONS
SEE PROFILE

SH Manjula
UVCE, Bangalore University
94 PUBLICATIONS **201** CITATIONS
SEE PROFILE

Venugopal K R
University Visvesvaraya College of Engineering
925 PUBLICATIONS **3,702** CITATIONS
SEE PROFILE

Lalit M Patnaik
Indian Institute of Science
837 PUBLICATIONS **8,674** CITATIONS
SEE PROFILE

International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

A Data Mining Perspective in Privacy Preserving Data Mining Systems

Kumaraswamy S[Á*], Manjula S H[Á], K R Venugopal[Á] and L M Patnaik[B]

[Á]Department of Computer Science and Engineering, Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001, India
[B]Indian Institute of Science, Bangalore, India

Accepted 20 March 2014, Available online 01 April 2014, **Vol.4, No.2 (April 2014)**

Abstract

Privacy Preserving Data Mining (PPDM) presents a novel framework for extracting and deriving information when the data is distributed amongst the multiple parties. The privacy preservation of data and the use of efficient data mining algorithms in PPDM systems is a major issue that exists. Most of the existing PPDM systems employ the cryptographic key exchange process and the key computation process accomplished by means of certain trusted server or a third party. To eliminate the key exchange and key computation overheads this paper discusses the Key Distribution-Less Privacy Preserving Data Mining (KDLPPDM) system. The novelty of the KDLPPDM system is that no data is published but only the association rules are published to achieve effective data mining results. The KDLPPDM embodies the C5.0 data mining algorithm for classification rule generation and data mining. The results discussed in this paper compare the C4.5 based KDLPPDM system with the C5.0 based KDLPPDM system and the efficiency in rule generation, overhead reduction and classification efficiency of the latter is proved.

Keywords: *Privacy preserving data mining, Association rules, C4.5 algorithm, See5.0/C5.0 algorithm, classification rules*

1. Introduction

Organizations and institutions like research establishments, business or corporate houses and government bodies possess certain framework or infrastructures for maintaining huge data collections for analyzing and processing it. The information extracted from its local or confined databases are not sufficient for accomplishing or facilitating the expected results. Therefore, such shortcomings do require a platform or system that will effectively collect the huge distributed data and can perform the data mining to get the expected information that will be analyzed efficiently and precisely. Such scenarios put forth the need for privacy preservation in data mining (*PPDM*) systems. The major objectives of the *PPDM* systems is to maintain the data integrity of the data published and to achieve efficient data mining results. For illustration there is the need for a highly secured system for preventing terrorism, the data collection and analysis of collected data from the immigration department is required to be done from every participating nation for analyzing the factions and track rebels or terrorists to counter any terror activities. Employing such kind of robust cooperative systems, the security agencies or intelligence bureau can effectively track the suspicious movements and can curb the possible misfortunes.

Another perspective use of such systems can be applied to monitor illegal immigrants. Meanwhile, such secured cooperative systems can effectively provide for facilitating data authorization and collaborative data utilization by numerous organization or independent entities without compromising with the integrity of data and its security. Such *PPDM* systems is employed for corporate houses too, where they are used for certain constructive and optimization activities.

1.1 Problem Formulation

In general, it is found that the entities involved in such system don't feel secured resulting in partial, inaccurate and doctored disclosure of information among every other party even if certain agreements and understanding are in place. Any organization considers privacy preservation as the dominant factor while dealing with such circumstances. A number of organizations are governed by certain rules and regulations that are required to be implemented for preserving the privacy of the information or data(V.S. Verykios, *et al*.,, 2004),(Y. Lindell and B. Pinkas, *et al*.,2000),(Yaping Li, *et al*., 2011), (O. Goldreich *et al*., 2002),(A.W.-C. Fu, *et al*., 2005) before it is released. In addition to the privacy concerns, there is a need for robust data mining mechanisms to be embodied in the *PPDM* systems to achieve the desired results and analysis from the data published by the parties involved. To summarise it can be stated that *PPDM* systems must provide for privacy preservation of the data published and

*Corresponding author: **Kumaraswamy S; L M Patnaik** is an Honorary Professor

effective data mining mechanism adoption to obtain the desired results.

1.2 Motivation

Most of the research work published in the area of $PPDM$ mainly concentrates on solving the issue of privacy preservation with less or negligible efforts drawn to enhance the data mining efficiency of the $PPDM$ systems. The $PPDM$ systems that currently exist secure the data using varied mechanisms. Cryptography is the most commonly used mechanism to achieve data integrity. To incorporate cryptographic techniques key generation and key exchange is an integral function to be achieved where in adversaries can benefit if improper techniques are adopted. To overcome this drawback and render the adversaries ineffective this paper introduces a Key Distribution-Less Privacy Preserving Data Mining ($KDLPPDM$) system in this paper. The adoption of the optimal data mining technique is also critical and must facilitate accurate analysis with the use of limited data. Limited work is carried out to study the effect of the varied data mining techniques in $PPDM$ systems which is a major motivating factor considered. Researchers have proposed to use data mining algorithms like $ID3$,$CART$,$C4.5$ algorithm in $PPDM$ systems. Limited work has been carried out considering the $C5.0$ data mining algorithm in $PPDM$ systems.

1.3 Contribution

To overcome the privacy preserving concerns and address the use of efficient data mining algorithms, in this paper the $KDLPPDM$ system is discussed. To preserve privacy the use of the $Commutative\ RSA$ cryptographic algorithm is considered. The $C5.0$ data mining algorithm is considered for mining in the $KDLPPDM$ system. The $KDLPPDM$ system discussed in this paper considers no key exchange to establish the $Commutative\ RSA$ algorithm which makes it robust even in the presence of adversaries. It is assumed that all the parties involved in $KDLPPDM$ system adhere to the semi honest trust model. The proposed $KDLPPDM$ overcomes the involved parties concerns of data publication by adopting the publication of the classification rules instead of the original data. The classification rule data to be published is secured by the $Commutative\ RSA$ cryptography mechanism and no rule data is exchanged in the original form there by addressing the privacy concerns of the parties involved. In the further section of the paper the construction of the secure rule set (i.e. the combined classification rules generated by all the parties) is discussed. To address the concerns relating to efficient data mining mechanisms adoption, the $C5.0$ data mining algorithm is used to overcome the shortcomings of the existing algorithms like $ID3$ and $C4.5$. It is to be noted that the proposed $KDLPPDM$ relies on the accuracy of the combined classification rules generated to achieve mining accuracy. To study the effect of the classification rules on the mining results the $C4.5$ based $KDLPPDM$ system is compared with the $C5.0$ based $KDLPPDM$ system in the experimental study presented in this paper.

1.4 Organization

The remaining manuscript has been organized as follows: The second section of the manuscript discusses few dominant related works or existing works while the section third presents the background of the research. Similarly, the ascending section (Section 4) discusses the $C5.0$ algorithm for data mining. Section 5 presents the system model for $KDLPPDM$ with all the required preliminary notations, system initialization, classification of rule sets and secure rule set generation. The Section 6 presents the Results and its discussion which is followed by Section 7 that presents conclusion of the proposed work and its future scopes.

2. Related Work – Privacy Preserving Data Mining

Initially the approach for data mining for unleashing its potential and benefits was started in early 1980's and its efficiency and effectives attracted business originations for its adaptation and other establishments a decade later (V.S. Verykios, *et al*.,, 2004), (Y. Lindell and B. Pinkas, *et al*.,, 2000) ignited the inception for provision of privacy preservation of the records by means of a fundamental PPDM framework that came into existence in the start of the 20th century. With its first illustration of robustness in data security a major of research society and organizations came ahead. A number of approaches and systems were proposed for PPDM requirements that will effectively secure the data and facilitate data mining. Unfortunately the overall or even the dominant objectives will not be accomplished and majority of researches do have certain limitations and performance constraints.

(Yaping Li *et al*.,, 2011) introduced and proposed for the categorization of PPDM technique into two dominant kinds, First Secure MultiParty Computation Techniques and Partial Information Hiding techniques.In their classification approach the authors' classified works on the basis of privacy level. In this work only the fundamental classifications had been done and the scopes for its optimization as per real time requirements are not addressed.

(O Goldreich *et al*.,, 2002),(A.W.C. Fu, *et al*.,, 2005)implemented a PPDM system that achieved a little higher level of data security and its privacy. In their work the authors proposed partial information hiding system with the implementation of secure multi-party computation technique. This system architecture is considered as optimum but its limitations for accuracy and computational complexity will not deliver the expectations of high data rate and higher user count environment. Even this work was functional with fundamental classification rules and data mining approaches which can't comply with huge requirements.

(D Agrawal, *et al*.,, 2001) discussed about an approach based on partial information hiding for PPDM application that illustrated confined privacy preservation characteristics in the occurrence of colluded users. The

partial information hiding approach was employed in Privacy Preservation in Data Mining systems which was further classified as data perturbation by (K Chen *et al*.,, 2005), (S. Papadimitriou *et al*.,, 2007).

(L Sweeney *et al*.,, 2002) proposed a scheme called k-anonymity approach which was further enhanced by (P.S Yu *et al*.,, 2004) and (Slava Kisilevich *et al*.,, 2010). Similarly, (W.DU *et al*.,, 2003) and (R. Agrawal *et al*., 2005)(http://www.rulequest.com/see5-info.html) introduced and worked for a retention replacement system. In their partial information hiding approach the data which is to be published for mining operations are transformed in such a way that in general doesn't influence the results of mining operations. This is the matter of fact that the partial information hiding approach is not competent for exhibiting performs in the occurrences of colluded users. Even this approach considers for locally generated rules as dominant but unfortunately, its functional efficiency will not be validated for real time multiple user environment. The limitations with these works were the consideration of performance cost and higher accuracy in real time scenario.

(Xindong Wu *et al*.,, 2008),(Tomasz Bujlow *et al*.,, 2012),(Po-Hsun. Sung *et al*.,, 2010) and (http://www.rulequest.com/see5-info.html) considered C5.0 data mining algorithm for data extraction and mining for exhibiting higher accuracy and efficiency. The researchers had implemented C5.0 algorithm and analyzed the performance of this system as compared to its predecessors with work done by (Xindong *et al*.,,2008) and (Ming-Jun *et al*., 2006) who employed ID3 and C4.5 respectively for accomplishing classification. This is the matter of fact that C4.5 exhibited higher accuracy and efficiency as compared to ID3 but considering few specific requirements like outfitting or adaptability of rules these approaches did not deliver significant results.

(C. Dwork *et al*.,, 2006) introduced an approach of data privacy while introducing noise that was further enhanced by (M. Islam *et al*.,, 2003) where the author implemented anonymization of the resource data present with the allied parties.

(Gregory J *et al*.,, 2011), (L. Sweeney,2002) and (P. S. Yu *et al*.,,2004) employed certain cryptographic based mechanisms which was considered by (Ming-Jun Xiao *et al*.,,2006), (J. Vaidya *et al*.,,2002),(Srikant *et al*.,,2000) and (B. Pinkas *et al*.,, 2002) where these research group considered to apply cryptographic approaches for facilitating privacy of the data. Since the work done by these researchers were depending on cryptographic approach, these all came out with certain accuracy enhancements but as the cost of computational cost and of course complexity. In case of cryptographic approaches, in the multi party computation, the extra overhead is added up sequentially and the key exchange approach makes the system costlier. Therefore these systems could not be considered as an optimum system. Even C5.0 algorithm is untouched in these approaches so, leaving behind a scope for further enhancement.

(Chris *et al*.,, 2002) and (Grandison *et al*.,, 2009) considered the uniqueness and the efficiency of the PPDM technique that not only provides security using commutative algorithms but also eliminates the overheads arising because of key distribution, key collection or storage overheads and re-keying process that was further considered in (Hubenko *et al*.,, 2007), (Kulkarni *et al*.,, 2011) and (Nathaniel *et al*.,, 2012). This system performed well but broad functions with robust classification and rule generation techniques like C5.0 was lacked and even commutative RSA kind of approaches were not presented in their work. Therefore, in order to explore the further enhancements and optimization here in this work the author has proposed a highly robust and efficient system called Key Distribution-Less Privacy Preserving Data Mining ($KDLPPDM$) that not only reduces the computational complexity and overheads but also enhances the privacy and accuracy

3. Background Work

(Piotr Andruszkiewicz *et al*.,, 2011) has presented robust system architecture for privacy preserving with few enhanced classification approaches and association rule mining on centralized kind of data sets. In their work they introduced the system for classification and association rule mining for privacy preservation of data sets which is distorted with randomization approach. In this approach the individual values or parameters are modified at random so as to facilitate an expected privacy preserving for private data. Here it has been considered that in overall operations of data deriving or extraction only the scattered or distorted and parametric attributes allied with the distortion procedure are revealed or disclosed for constructing a classifier and developing association rules for data mining. The authors proposed an approach for MMASK optimization which is the enhanced form of MASK algorithm, for eliminating the exponential complexity of an attribute for estimation in related with cardinality. The developed protocol provides individual attribute to posses its own distortion parameters. In this work the authors illustrated how to employ the randomization approach for integer as well as ordinal attributes so as to alter their respective values for obtaining same distribution of values after distribution. This made the system to perform higher accuracy with enhanced privacy. Additionally the authors implemented a privacy preserving scheme for performing classification on the basis of emerging patterns and have provided privacy preserving modifications like eager ePPCwEP and CAEP and similarly lazy lPPCwEP and DeEPs classifier. In spite they have employed procedures of bagging as well as boosting algorithms with reconstruction of decision trees that comes out to be with higher accuracy of classification. In their work the author illustrated that meta-learning facilitates higher accuracy gain for applications like in privacy preserving classification.

This is the matter of fact that this work has the higher span for privacy preserving approach but still this system will not be stated as an optimum solution. The presented work employs the priory based classification which ultimately lacks in ultimate optimized results generation and on the other hand the computational complexity is much higher and therefore can't be considered for real

time general applications. On the other hand the proposed mechanism covers a broad area but ignore data mining algorithms like C5.0 algorithm. Even it compares result with ID3 algorithm that is predecessor of highly robust C5.0 algorithm. Therefore this work lacks in few regions which can be eliminated without mammoth task and huge complications. Even this work doesn't emphasize on privacy aspects. Our proposed system considers C5.0 algorithm which performs better as compared to this reference work in every performance and functional criteria.

4. C 5.0Data mining Algorithm

*4.1 Decision Tree Learning Algorithms and C*5.0

C5.0 is a highly robust and efficient successor of C4.5 algorithm that provides higher accuracy and least memory occupancy with maximum classification preciseness.

Dominantly the algorithm C5.0 is implemented for creating rule sets which is further employed for classifying data samples or records. The characteristics like boosting and weighting are also supported in the C5.0 algorithm.

The amount of information present in certain attributes that is referred as $Entropy$, plays a significant role in deciding the effectiveness and performance of decision tree learning in data mining. Mathematically the entropy of a random variable x (according to information theory) is given as:

$$H(x) = \sum_x p(a)\, log_2 \frac{1}{p(a)},$$

Where $'a'$ refers for attributes and $p(a)$presents value of class attributes. Similarly the conditional entropy can also be presented for *'a'* provided another attribute $'y'$.

$$H(x/y) = \sum_{x,y} p(x,y)\, log_2 \frac{1}{p(x,y)}$$

for x, i.e. $H(x/y) \leq H(x)$.

In case of uncertainty suck kind of reduction is referred as mutual information between attributes and given as

$$I(x;y) = H(x) - H(x/y)$$

Initially, in order to make individual partition the y_i variable that could facilitate the maximized information about another x variable the information gain I(x; y)is optimized. Since, this phenomenon encompasses of implementation of a parallel scheme for assisting y_i with multiple outputs, the C5.0 algorithm optimizes the relation that gives the Gain ratio and is defined as

$$I(x;y_i) \Big/ H(y_i)$$

In addition, for avoiding the selection lower value of entropy by certain attribute, the Gain ratio is enhanced and for that the robust C5.0 algorithm performs adaptive weighting by increasing value of $H(y_i)$.

A general issue occurred in majority of decision tree generation, classification or rule generation approaches is the adaptability of training sets, normally referred as outfitting. In order to eliminate suck kind of problems and to enhance the existing approaches the C5.0 algorithm introduces an approach called Pruning. This mechanism has been presented in subsequent section.In highly robust system architectures and processing for secure competitive multi-party communication environment C5.0 facilitates a highly optimized solution for rule generations as well as classification. Few of the dominant features of C5.0 are like adaptive boosting, pruning and rule set generation. This section of the presented paper briefs about the C5.0 features and the contribution of its robustness for achieving optimum privacy preservation in data mining.

4.2 Adaptive Boosting

One of the significant characteristics of C5.0 algorithm is Adaptive boosting. The dominant idea behind processing adaptive boosting is to generate numerous classifiers on the information or the training data available with individual parties. Whenever an unauthenticated or unique sample or data is encountered for classification, the predicted class of the sample or data encountered is a weighted count of votes from individually trained classifiers. The data mining algorithm C5.0 generates numerous classifiers by initializing a single classifier. The created classifier in its ascending phase is developed by performing re-training on the data samples employed for creating the initial or first classifier, but being highly cautious to the cases of the training data set where the first classifier has classified incorrectly. These all results into the generation of a different data set by second classifiers than first classifier. The predominant algorithmic concepts behind the adaptive boosting approach have been given as follows:

- Select $\mathcal{K}$ samples from the available data sets, where individual being provided a probability of $1/N$ to train a classifier.
- Perform classification process on the data sets available with each party with the trained classifier.
- Substitute the samples or data available by multiplying the probability of the wrongly-classified samples by a weight B.
- Now, continue previous steps n times with the generated probabilities.
- Mingle the A classifiers by providing a weight value $log(BA)$ on individually trained classifier.

In this process the adaptive boosting process is invoked by means of C5.0 algorithm as well as the number of classifiers generated.

4.3 Pruning Process

In general the proposed algorithm C5.0 generates the decision trees in two consecutive steps. In the first step it constructs or generates a classifier which is suitable with the available data sets and then it performs pruning on the classifiers generated so as to avoid the possibility of over-

fitting of the data available with individual party. In the overall phenomenon two dominant options might be employed for affecting the approach in which the decision trees have been pruned. The initial option characterizes the degree in which the decision tree can make itself compatible with the training data and for this it characterize the minimum number of data sets that can pursue minimum of two branches at any of the available or created nodes in the decision tree. In fact this step aims to prevent the over-fitting of data by means of stopping training process until it get over-fitted with data sets. The other pruning option of C5.0 algorithm is the security feature that influences the algorithm by means of post-pruning of constructed decision trees and rule sets generated. This overall pruning process is exhibited by eliminating certain parts of the generated decision tress or rule set generated which do possess a relatively higher error rate on data sets available on individual parties.

After eliminating the issue of outfitting by means of pruning approach the robust data mining algorithm C5.0 performs a post-pruning phase that replaces a branch of the tree by a leaf when the count of predicted errors for the last step is lower as compared to one for the branch. Thus it facilitates the accurate results with minimum computational complexity.

4.4 Rule Sets Generation

The algorithm C5.0 functions optimally for conversion of decision trees into rule sets. Here in proposed KDLPPDM PPDM model C5.0 algorithm has been used for generating rule sets for classification. Here the strength is that the generated rule sets are flexible for understanding as compared to decision tress and this approach can easily be implemented in terms of computational complexity and efficiency.

5. Key Distribution-Less Privacy Preserving Data Mining (*KDLPPDM*) System

In this section of the presented manuscript the proposed KDLPPDM system model is discussed. The presented PPDM protocol considers a multiple stages based (here 3 steps) function along with the system or protocol initialization phase. In this proposed system model for initializing the system, individual beneficiaries or parties do generate mining rules locally by adopting the C5.0 algorithm on the data available, which has already been pre-classified. The system initialization phase discusses the C5.0 algorithms along with it optimistic implementation with Commutative RSA based element based score matrix implementation for privacy preservation.

The implementation phases in the proposed $KDLPPDM$ protocol are as mentioned below:

1) The facilities of the data privacy and integrity of the rules generated locally and the development of the secure combined rule sets. This secure combined rule set is nothing else but the combination of all the rules generated locally by individual beneficiaries or parties in the form of encrypted data to accomplishing the privacy preservation goal.
2) After getting the privacy factor in the rules sets generated locally and developing the secure combined rule sets, in the further step these are employed for data mining and its analysis. In this phase an individual party or beneficiary is initialized which can propagate the secure rule sets across the network and among all the parties. Now, the beneficiaries or the parties decrypt the rules achieved form initiator by employing decryption approach. Here the author proposes for $Commutative\ RSA$ enabled data mining approach for $PPDM$ requirement. The initiator which does decrypts the secure rule sets ultimately accomplishes the combined rule set creation.
3) In the ascending phase of the proposed or developed system model the results obtained from mining which is unclassified till, is processed with the $C5.0$ classification algorithm which ultimately gives precise and accurate classification with highly efficient results.

Now in the ascending section, we will be discussing the dominant preliminary notations employed in the system formulation, development and description of the model. The list of notations used throughout this paper is summarized in Table 1 given below

Table 1: List of Notations used

$\mathcal{usr}$	The combination of parties or users involved in the $PPDM$ protocol of framework.
$g_{ph}^{c5.0}(Ut, Rl_{Ut})$	$C5.0$ classification function
Ut	states $C5.0$ based classified or partitioned data available with all the users or parties
$Et_{n\mathcal{R}t}$	data present with individual party of the user set $\mathcal{usr}$
$T\mathcal{R}_{n\mathcal{R}}$	Transactions made for $\mathcal{R}$ attribute set
Ph_r	Classification set
$g_{\mathcal{R}hgen}^{c5.0}(Ut, Ph_{\mathcal{R}}) = Nh$	rule set constructed locally on the basis of the data Ut
$\mathcal{SR}_{set}$	Secure rule set generated
N_{Set}	Combined rule set generated
UUt_n	Set of unclassified data
$g_{\mathcal{R}hm\mathcal{RG}}^{C5.0}$	function for overall combined rule sets generation

5.1Dominant preliminary notations employed in $KDLPPDM$

Consider a set of users or parties is presented by $\mathcal{Usr}$ that refers n users involved in the PPDM model. Here it can be considered that the all n users are agree for participating in a Semi − Honest Trust Model. The user set $\mathcal{Usr}$ can be presented by

$$\mathcal{Usr} = \{usr_1, usr_2, usr_3, \dots\dots\dots usr_n\}$$

In other way $\mathcal{U}sr$ can be presented as

$$\mathcal{U}sr = \{usr_1, usr_2, usr_3, \dots \dots \dots usr_m, usr_n\}$$
Where $m \neq n$ and $1 \leq m <$ n

Consider, usr_{rst} refers a set of users involved in privacy preservation in data mining model at while excluding party usr_n, is given as

$$usr_{rst} = usr \cap usr_n$$
$$usr_{rst} = \{usr_1, usr_2, \dots usr_m, usr_n\} \cap usr_n$$
$$usr_{rst} = \{usr_1, usr_2, usr_3, \dots \dots \dots usr_m\}.$$

Consider, the set Ut which represents the vertically portioned data available with all the n users or beneficiaries in the proposed privacy preservation model which is mathematically given as

$$Ut = \{Ut_1 \cup Ut_2 \cup Ut_3 \dots \dots \cup Ut_n\}$$

In the proposed $PPDM$ model it has been assumed that the individual party or user encompasses its classified and unclassified data. Here the data Ut encompasses the $T\mathcal{R}$ transactions that it self consists of Et attributes. In the data available amongst the individual n parties Ut consisting of $T\mathcal{R}$ sets of transactions are assumed to be have no similar attributes amongst themselves. The data set present with the n^{th} user or party can be given as follows:

$$Ut_n = \{T\mathcal{R}_{n1}, T\mathcal{R}_{n2}, \dots T\mathcal{R}_{n\mathcal{R}}\} \text{ and } Ut_n \neq \emptyset$$

In above expression, the variable $\mathcal{R}$ presents the total count of transactions made where individual transaction $t\mathcal{R}_{nr}$ encompasses t exclusive attributes states as

$$T\mathcal{R}_{n\mathcal{R}} = \{Et_{n\mathcal{R}1}, Et_{n\mathcal{R}2}, \dots Et_{n\mathcal{R}t}\}$$

The data present with the n^{th} user or the beneficiary party $\mathcal{S}$ given by the expression

$$Ut_n = \begin{Bmatrix} \{Et_{n11}, Et_{n12}, \dots Et_{n1t}\}, \{Et_{n21}, , \dots Et_{n2t}\}, \dots \\ .\{Et_{n\mathcal{R}1}, Et_{n\mathcal{R}2}, \dots Et_{n\mathcal{R}t}\} \end{Bmatrix}$$

In the above presented expression the variable set $Ut_n \subset Ut$.

Since $\mathrm{Et}_{n\mathcal{R}t} \in \mathrm{T}\mathcal{R}_{n\mathcal{R}}$ and $\mathrm{T}\mathcal{R}_{n\mathcal{R}} \in \mathrm{Ut_n}$ it can be stated that $\mathrm{Et_{nrt}} \in \mathrm{Ut_n}$ and $\mathrm{Et}_{n\mathcal{R}t} \notin \mathrm{Ut} \cap \mathrm{Ut}_n$ as the data present with individual party of the user set $\mathcal{U}sr$ is processed with C5.0 algorithm. Ut Refers the pre classified set of data $\mathrm{Ut_n}$ which is present with individual parties and follows $usr_n \in \mathcal{U}sr$. Individual users or party usr_n encompasses a data set UUt_n representing an unclassified set of data that is required to be classified. Since, the rules generated locally facilitates lower accuracy in data mining proved by the research presented in (Patrick Sharkey *et al*.,, 2008) therefore the users involved in the set $\mathcal{U}sr$ consent for sharing the rules generated locally for accomplishing a higher degree of accuracy in mining in the case of semi-honest trust model (Yaping Li *et al*.,, 2012).

5.2 KDLPPDM System Initialization

The proposed privacy preservation approach mentioned in this manuscript considers that the available dataset Ut is partitioned by means of vertically partitioning and the data Ut_n which has been pre classified with the usr_n is employed for generating the classification rules by the C5.0 data mining algorithm. The higher classification accuracy and robustness of the C5.0 algorithm makes is to be considered in spite of other existing approaches and classification tree algorithms such as ID3(C. Dwork *et al*.,,2006),(Xindong Wu *et al*.,, 2008) and C4.5(Ming-Jun Xiao *et al*.,, 2006),(M. Islam *et al*.,, f2003).

Consider $\mathrm{g}_{\mathrm{ph}}^{\mathrm{c5.0}}(\mathrm{Ut}, \mathrm{Rl}_{\mathrm{Ut}})$ defines the C5.0 classification function and data set $\mathrm{Ut} \in \mathrm{Ut_n}$. The variable $\mathrm{Rl_{dt}}$ refers for the classification rules achieved from the available data Ut. The function for generating rules is based on the pre classified data sets Ut as well as the classification set $\mathrm{Ph}_{\mathcal{R}} = \{\mathrm{ph}_1, \mathrm{ph}_2, \dots \mathrm{ph_p}\}$ can be presented by expression

$$\mathrm{g}_{\mathcal{R}\mathrm{hgen}}^{\mathrm{c5.0}}(\mathrm{Ut}, \mathrm{Ph}_{\mathcal{R}}) = \mathrm{Nh}$$

Where Nh refers the rule set constructed on the basis of the data Ut and of course on the classification set $Ph_{\mathcal{R}}$.

The classification rule set Nh generated locally are shared or employed by all of the participating parties or entities of the set $\mathcal{U}sr$ for performing analysis. The preservation of privacy for the locally generated rule set is accomplished by taking into consideration of Commutative RSA security approach. This is required to generate the preserving facilities for multiple parties. The research phases for system initialization will be effectively understood and implemented with the help of the following algorithms: Once the initialization phase is completed the individual parties usr_{n} of the set $\mathcal{U}sr$ carries its security constraints for preserving the privacy of the rules $\mathrm{Nh_n}$ which has been generated by means of proposed C5.0 data mining algorithm. Those rules which are generated locally are employed for the development of secure rule sets by implementing Commutative RSA based privacy preserving. The predominant purpose of the proposed privacy preservation approach or PPDM system is to provide or construction of the combined rule sets $\mathrm{N_{Set}}$ which is defined by the privacy preservation. In expression it is given as follows:

$$N_{Set} = \{Nh_1 \cup Nh_2 \cup Nh_3 \dots . \cup Nh_n\} \forall n$$

In the above mentioned expression the variable n presents the beneficiary or party $usr_n \in \mathcal{U}sr$. The combined rule set $\mathrm{N_{Set}}$ is employed for attaining the mining consequences on the unclassified data set presented as UUt_n.

Table 2: Algorithm for KDLPPDM system initialization

Algorithm Name : *KDLPPDM* Initialization

Input **:** Two prime numbers *P* and *x* where $P > x$

Output::Encryption key=(n_{usr_n} , a_{usr_n}), Decryption key = (n_{usr_n} , D_{usr_n}) Local classification rules, Nh_n *for each party*

1. *Require 2 prime numbers P* and *x* where $P > x$
2. ***For each*** *party* $usr_n \in \mathcal{U}sr$
3. ***Initialize*** *parameters of* P_{usr_n} *Privacy Preserving Function*
4. $P_{usr_n} = P$
5. $x_{usr_n} = x$
6. *Compute* $n_{usr_n} = P_{usr_n} \times x_{usr_n}$
7. *Compute* $\varphi_{usr_n} = \varphi\left(p_{usr_n}\right) \times \varphi(q_{usr_n}) = \left(P_{usr_n} - 1\right) \times (x_{usr_n} - 1)$
8. *Obtain* $a_{usr_n} \mid GCD\left(a_{usr_n}, x_{usr_n}\right) = 1$
9. *Obtain* $d_{usr_n} = {e_{usr_n}}^{-1}\, \mathcal{M}od(\emptyset_{usr_n})$
10. *Party* P_{usr_n} *encryption key* = (n_{usr_n} , a_{usr_n})
11. *Party* P_{usr_n} *decryption key* = (n_{usr_n} , d_{usr_n})
12. ***End Initialize*** *parameters of* P_{usr_n} *Privacy Preserving Function*
13. ***Initiate Rule Generation*** *of* $g^{c5.0}_{\mathcal{R}hgen}(Ut, Ph_{\mathcal{R}})$ *(the procedure for rule generation using C5.0 has been presented in Sub section 5.2)*
14. *Obtain Pre classified data set of* $usr_n Ut_n$
15. *Obtain Classification Set* $Ph_{\mathcal{R}}$
16. *Compute local classification rules* $g^{c5.0}_{\mathcal{R}hgen}(Ut_n, Ph_{\mathcal{R}}) = Nh_n$
17. ***End Rule Generation***
18. ***End for each.***

In order to perform the analysis purpose, all the rules Nh_n are generated locally from individual party. It must be realized that the factual data Nh_npreset with party n is not negotiated and it is not shared amongst the numerous users for providing privacy like in the customary models (Pui K. Fong *et al*.,, 2012). The overheads created in data mining protocol and its initialization can be presented as follows:

$$T_{Init} = T_b + T_{dp}$$

In the above presented expression the variable T states for the computational overhead realized in terms of the duration of computation and T_b states the computational overhead relevant in the estimation of parameters like n , φ and the $\mathcal{GCD}$. The variable T_{dp} indicates the overheads created due to unwanted encryption and decryption and even huge computation. Consider that $B(n)$ presents the function of time complexity engrossed in performing n bit operations.In the initialization process the overall computational overhead created can be presented as follows:

$$T_{Init} = B(2\mathcal{P}) + B(2\mathcal{P}d^2_{\mathcal{P}e}) + B(\mathcal{P}d^2_{\mathcal{P}e})$$

5.3 Step 1: Provisioning of Privacy Preservation of the Classification Rules and the generation of Secure Rule Sets

The initial phase or the step of the proposed privacy preservation for data mining approach is emphasized for creating the secure rule sets given as SN_{Set} and the characteristics of privacy preserving of the locally generated rule sets. In order to facilitate the privacy facility for individual user the involved users must have to releases its locally generated rules Nh_n only after processing it for preservation or key assignment (n_{usr_n} , a_{usr_n}). The other participants also perform for privacy preservation or certain key provisioning for its rules Nh_n by means of employing its allied respective security parameters or keys. The function for commutative encryption of certain data A by implementing encryption key (n, a) can be presented as follows:

$$\mathcal{E}\,(A) = A^a\, \mathcal{M}od(n)$$

Now, enhancing or elaborating the definition presented above in a form of a multi party set-up the function for encryption which is generally executed by certain user or party n can be presented as follows:

$$\mathcal{E}_n(\mathrm{A}) = \mathrm{A}^{\mathrm{a}_{usr_n}}\, \mathrm{Mod}(n)$$

Consider the variable $\mathrm{SN_{Set}}$ presents the secure rule set which has to be constructed. Then the secure rule set can be given as:

$$SN_{Set} = \{SNh_1 \cup SNh_2 \cup SNh_3 \ldots . \cup SNh_n\}\, \forall\, n$$

In the above presented expression the variable n gives the total count of parties considered in the privacy preserving system model and SNh_n states the secured rule set of party usr_nachieved after n number of encryption processes. In the presented work the algorithm employed for providing privacy preserving for the rule sets generated, has been given as follows: Considering the above presented algorithm it can be found that the rules present with individual party or user are in general encrypted for n times. In order to develop the secure rule sets there subsist a communication cost owing

Table 3: Algorithm for Secure rule-set generation

Algorithm Name : Secure Rule Sets Generation SN_{Set}

Input: Encryption key = (n_{usr_n} , a_{usr_n}) ,Local classification rules $Nh_n$of each party

Output: Secure Rule Set SN_{Set}

1. Initialize $SN_{Set} = \emptyset$
2. **For each** $usr_n \in usr$
3. Compute $SNh_n = \mathcal{E}_n(Nh_n) = {Nh_n}^{a_{usr_n}}\, Mod(n)$
4. Compute $usr_{rst} = usr \cap usr_n = \{usr_1, usr_2, usr_3, \ldots \ldots \ldots usr_m\}$ where $m \neq n$
5. **For Each** $usr_m \in usr_{rst}$
6. Compute $SNh_{mn} = \mathcal{E}_{mn}(SNh_n) = {SNh_n}^{e_{usrm}}\, Mod(n)$
7. **End For**
8. $SN_{Set} = SN_{Set} \cup SNh_{123\ldots mn}$
9. **End For each**

to the relocation of the secure rules across the framework or among all participating users.

Now, the cost incurred in communication $Comm_{Step1}$ for the initial step can be given as follows:

$$Comm_{Step1} = B(2(n-1)\mathcal{F}).$$

In the above mentioned expression the variable $B(n)$indicates towards the communication function n representing all those participants or parties which are involved and the total amount of data bit transacted is given as $\mathcal{F}$. It represents a vital step functional since this phase facilitates the privacy of the data sets or data by means of robust commutative cryptography systems. In the circumstance of a malicious user or participants functional in the proposed system can be developed in such way that even if getting any data, the achieved data remains in a cryptic format with zero knowledge so that it can provide privacy whether how many time encryption is done. Then while, the information transacted among the participants is computationally impossible to differentiate. In order to have a better understanding consider that the involved parties are presented by $\mathbb{P}, \mathbb{Q}\ and\ \mathcal{N}$and $(\mathcal{W}_p, U_p), (\mathcal{W}_x, U_x)$ and $(\mathcal{W}_{\mathcal{R}}, U_{\mathcal{R}})$ states functions for their encryption and decryption correspondingly. Consider the data present with variable $\mathbb{P}\ and\ \mathbb{Q}$ is given by A which is available for providing N. In order to get the provisioning or privacy the involved parties $\mathbb{P}\ and\ \mathbb{Q}$ facilitate the data to N after performing encryption on the data with their respective encryption functions given as $\mathcal{W}_{\mathcal{P}}(A) and\ \mathcal{W}_x(A)$. The facility of privacy is given if $\mathcal{W}_{\mathcal{P}}(A) and\ \mathcal{W}_x(A)$ is retrieved by N are computationally indistinguishable. Mathematically, $\mathcal{W}_{\mathcal{P}}(A) \neq \mathcal{W}_x(A)$. The computational indistinguishablity of the proposed PPDM model has been presented in the ascending section of the manuscript.

5.4 Step 2: Generation of Combined Rule Sets while preserving data privacy

In this phase of PPDM system modeling it is tried to accentuate for generating the combined rule sets after generating secure rules as done in previous section. The secure rule sets SN_{Set} is generated that encompasses the classification rules of all the participating individuals $usr_n \in \mathcal{U}sr$ encrypted n times. Consider an individual party $usr_{int} \in \mathcal{U}sr$ referring the initiator which wants to generate the combined rule sets. The initiator usr_{int} proliferates the secure rule sets SN_{Set}among all the participants that decrypt the complete secure rule sets achieved or generated and propagate it back to the other users or parties. Consider usr_{rst} state the participants set excluding the initiator usr_{int} . mathematically, it can be given as

$$usr_{rst} = \mathcal{U}sr \cap usr_{int} \text{Where } usr_{int} \in usr$$
$$usr_{rst} = \{usr_1, usr_2, usr_3, \dots\dots\dots usr_m, usr_{int}\} \cap usr_{int}$$
$$usr_{rst} = \{usr_1, usr_2, usr_3, \dots\dots\dots usr_m\} \quad \text{Where } m \neq int$$

The decryption of the secure rule sets at individual set is accomplished by means of Commutative RSA approach with respective keys(n, D).The function employed for performing commutative decryption process has been given as follows:

$$U(A) = A^D Mod(n)$$

Now, elaborating the mentioned definitions to a multi-party communication environment for a participants or beneficiary $n$can be presented as:

$$U_n(A) = A^{D_{usr_n}} Mod(n)$$

Consider Nh_n depicts the $C5.0$ classification rules of the individual beneficiary or participants in PPDM $usr_n \in usr$. Thus, the combined rule sets can be presented as

$$N_{Set} = \{Nh_1 \cup Nh_2 \cup Nh_3 \dots . \cup Nh_n\}$$

The modeled algorithm being implemented in the generation or development of the combined rule sets can be given as follows:

Considering the above presented algorithm it can be found that the initiator usr_{int} proliferates the secure rule sets SN_{Set} for all the participants usr_{rst} bound in the proposed $KDLPPDM$ system. Individual parties perform the process of decryption of the secure rules generated and then it is sent to the ascending or next participants with such manner that the party initializing the steps performs decryption at last. Then the initiator party achieves the secure rule set decrypted precisely $(n-1)$ times. The last and of course ultimate decryption of the secure rule set which is exhibited by the initiator facilitate the combined rule set given as N_{Set}.

Table 4: Algorithm for combined rule-set generation

Algorithm Name : Combined Rule Set Construction N_{Set}

Input: Decryption key = (n_{usr_n}, D_{usr_n})of each party and Secure Rule Set SN_{Set}
Output: Combined Rule Set N_{Set}

1. Initialize $N_{Set} = \emptyset$
2. Initialize initiator $usr_{int} \in usr$
3. Compute $usr_{rst} = usr \cap usr_{int} = \{usr_1, usr_2, usr_3, \dots\dots\dots usr_m\}$ where $m \neq int$
4. **For each** $U_m \in usr_{rst}$
5. Initialize $Temp_m\ N_{Set} = \emptyset$
6. **For each secure rule** $SNh_n \in SN_{Set}$
7. Compute $D_m(SNh_n) = SNh_n^{D_{usrm}} Mod(n)$
8. $Temp_m N_{Set} = Temp_m N_{Set} \cup D_m(SNh_n)$
9. **End For**
10. $SN_{Set} = Temp_m N_{Set}$
11. **End For**
12. Initialize $Temp_{init}\ N_{Set} = \emptyset$
13. **For each secure rule** $SNh_n \in SN_{Set}$
14. Compute $D_{init}(SNh_n) = SNh_n^{D_{usrinit}} Mod(n)$
15. $Temp_{init} N_{Set} = Temp_{init} N_{Set} \cup D_{init}(SNh_n)$
16. **End For**
17. $\mathbf{N_{Set} = Temp_{init}\ N_{Set}}$

The communication cost $Comm_{Step2}$ for the presented first phase can be given as follows:

$$Comm_{Step2} = B(2(n-1)\,\mathcal{F})$$

In the above presented expression $B(n)$states communication function while the variable n gives the parties participating and $\mathcal{F}$ refers the total data bits transacted. It can be expressed as the bits transferred.

$$\mathcal{F} \propto SN_{Set}.$$

Although, this development phase is efficient for constructing the combined rule sets N_{pool} in secured way without any fear of privacy factor since the transacted data is always in the form of cryptic thus facilitating security features in the presence of malicious participants. Furthermore all the participants involved are in general aware of the security features like encryption and decryption keys being employed in the proposed $KDLPPDM$ system model as it doesn't have any process of key exchange.

5.5 Step 3: Data Analysis with $\boldsymbol{C5.0}$ *Algorithm employing Combined Rule Sets*

In the presented paper, the robustness as well as effectiveness of $C5.0$ data mining algorithm has already been discussed. Here in this section a brief of $C5.0$ algorithm for association rule mining has been presented.

Association Rule Mining Algorithms

In the process of data mining the generation of association rule refers towards investigating the inter-relationships among numerous data sets available in data sets of the individual parties. Here in this paper and the proposed $KDLPPDM$ model the expected results generating from association rules states the relationships amongst the data sets of participants or parties and demographic information etc.

In general the association rules are given in the form of $X => Y$. Here the variable X and Y represents the two exclusive subsets for those all variables. In the proposed approach the association rule mining has been divided into dominant two parts, the first one represents the generation of regular item set which is succeeded by application of interestingness measurement for generating ultimate rules. In general the support rate is the dominant one which is employed for generating regular data sets.

Analysis of data available with C5.0 Algorithm

As discussed in earlier sections, the initial two phases of the proposed $KDLPPDM$ system model emphasizes on the facilitation of a highly robust system for achieving privacy of the exchange of classification rules. This phase presents the algorithms considered for performing analysis of the raw or unclassified datasets present with the individual users or parties. As soon as the individual generated rules are collected, then these rules are fed for combining which results into the generation of a highly efficient and optimum classification rule that can optimize the data mining operations.

Consider a variable UUt_npresenting the set of unclassified data present with parties $usr_n \in \mathcal{U}sr$ and the combined rule generation with n accumulated rules from individual party), N_{Set} is presented as

$$N_n^{C5.0} = g_{\mathcal{R}hm\mathcal{RG}}^{C5.0}(N_{Set}, n) \qquad \text{In other way,}$$
$$N_n^{C5.0} = g_{\mathcal{R}hm\mathcal{RG}}^{C5.0}(\{Nh_1 \cup Nh_2 \cup Nh_3 \ldots \cup Nh_n\}, n)$$
$$N_n^{C5.0} = g_{\mathcal{R}hm\mathcal{RG}}^{C5.0}\big(\{\{\mathcal{R}h_1, \mathcal{R}h_{21} \ldots, \mathcal{R}h_{1musr1}\} \cup \{\mathcal{R}h_{12}, \mathcal{R}h_{22} \ldots, \mathcal{R}h_{1musr2}\} \cup \ldots. \cup \{\mathcal{R}h_{1n}, \mathcal{R}h_{2n} \ldots, \mathcal{R}h_{1musrn}\}\}, n\big)$$

Now, considering these all expressions it can be found that the overall combined rule sets $N_n^{C5.0}$comprised of all the rules generated by usr_1 to usr_n. The highest possible number of generated rules by certain party or user usr_n can be given as $\mathcal{R}h_{1musrn}$.

The variable $g_{\mathcal{R}hm\mathcal{RG}}^{C5.0}$depicts the function for overall combined rule sets generation.

Consider the function for performing classification by means of $C5.0$ data mining algorithm is given by

$$Ph\_U\mathcal{M}\_Nh\mathcal{S}t_n^{C5.0} = g_{clss}^{c5.0}(N_n^{C5.0}, UUt_n)$$

In the above presented expression the variable $g_{clss}^{c5.0}$states the function for C5.0 classification that takes into account of two dominant factors. First the classification rules

Table 6: Datasets Considered For Evaluation

ALGORITHM NAME						C5.0 ALGORITHM		C4.5 ALGORITHM	
DATASET ID	DATA SET	NO OF CLASSES	NO OF ATTRIBUTES	NO OF TRAINING RECORDS	NO OF TESTING RECORDS	DECISION TREE SIZE	N_{Set} SIZE	DECISION TREE SIZE	N_{Set} SIZE
1	BREAST CANCER DATA (UNIVERSITY OF WISCONSIN)	2	10	699	699	16	11	49	16
2	SPAM E-MAIL DATABASE	2	57	4601	4601	73	133	393	163
3	1984 UNITED STATES CONGRESSIONAL VOTING RECORDS DATABASE	2	16	300	135	2	9	25	13
4	LANDSAT MULTI-SPECTRAL SCANNER IMAGE DATA	6	36	4435	2000	186	151	505	284
5	CREDIT APPROVAL DATA SET	2	15	490	200	22	26	89	34

Table 5: Algorithm for combined rule-set generation

Algorithm Name : C5.0 Mining Algorithm for data analysis and mining

Input: Overall rule sets generated N_{Set}
Output: Results obtained after data mining and analysis $Ph\_U\mathcal{M}\_Nh\mathcal{S}t_n^{C5.0}$

1. Initialize the combined rule set $N_n^{C5.0} = \emptyset$
2. Initialize $N_{pnt} = 0$
3. **For each** $Nh_n \in N_{Set}$
4. **For each** $\mathcal{R}h_{1musrn} \in Nh_n$
5. $N_{pnt} = N_{pnt} + 1$
6. $N_{ntmp}^{c5.0} = g_{\mathcal{R}hm\mathcal{R}g}^{C5.0}(\mathcal{R}h_{1musrn}, n)$
7. **End For**
8. $N_n^{c5.0} = N_n^{C5.0} + N_{ntmp}^{C5.0}$
9. **End For**
10. Calculate $Ph\_U\mathcal{M}\_Nhst_n^{C5.0} = g_{clss}^{C5.0}(N_n^{C5.0}, UUt_n)$

$N_n^{P5.0}$and the second those data sets UUt_n is present with user n as inputs. The output of the function is denoted as $Ph\_U\mathcal{M}\_Nh\mathcal{S}t_n^{C5.0}$.

The respective algorithm for performing data analysis and mining in proposed $KDLPPDM$ system model has been given as follows:

In order to accomplish the goal of analyzing the unclassified data UUt_n the classification function $Ph\_U\mathcal{M}\_Nhst_n^{C5.0}$ is employed. The presented system model $KDLPPDM$ facilitates an algorithm for preserving the privacy of the critical data shared for utilization in a multi-party communication environment and it facilitates a rule based classification analysis approach while employing $C5.0$algorithm that establishes itself as the best and optimum scheme for decision tree generation (http://www.rulequest.com/see5-info.html),(Xindong Wu *et al*.,, 2008),(R. Agrawal *et al*.,, 2000) . The use of the $C5.0$ algorithm over the existing $C4.5$ decision tree algorithm is justified based on the results presented in the next section of this paper.

6. Results and Analysis

The major goal of the n parties to come together in the proposed KDLPPDM system is to achieve better data mining results so as to establish useful analysis and conclusions from the data stationed in them. Hence it can be stated that the acceptance of the proposed KDLPPDM system heavily relies on the accuracy of the data mining algorithm adopted in it. In this section of the paper the $KDLPPDM$ system is developed with two decision tree algorithms namely the $C4.5$ and $C5.0$, and the results obtained are compared and presented here. The environments for simulations are maintained uniform for all the scenarios presented here.

The proposed KDLPPDM was developed using C# and C++ as the programming languages. The data mining algorithms i.e. the $C4.5$ and the $C5.0$ algorithms were run on the GCC complier on the Linux platform. Varied datasets have been considered to evaluate the performance and the details are summarized in the Table 6 given above.

The KDLPPDM was developed for a 3 user scenario i.e. $\mathcal{U}sr = \{usr_1, usr_2, usr_3\}$. The initialization and the tri-step approach are conducted in accordance to the details discussed in the previous section of this paper. The data mining efficiency of classification is dependent on the rules generated using both the data mining algorithm and is evaluated on the basis of the Receiver Operating Characteristics (ROC). The secure rule sets constructed N_{Set} are evaluated on the training and the test data sets. The ROC curves are studied and presented in this paper.

For $Data\ set\ ID = 1$ the number of training records and the number of testing records are equal so the ROC curve of the training and testing are identical and is shown in Fig. 1. of this paper. In the total of the 699 records the $C5.0$ misclassified 11 records when compared to the 16 records misclassified by the $C4.5$ data mining algorithm. The time required for classification was found to be 0.1 seconds and 0.46 seconds for the $C5.0$ and $C4.5$ data mining engines. From Fig.1 the area covered under the ROC curve of $C5.0$ is 0.98 and of $C4.5$ is 0.96.

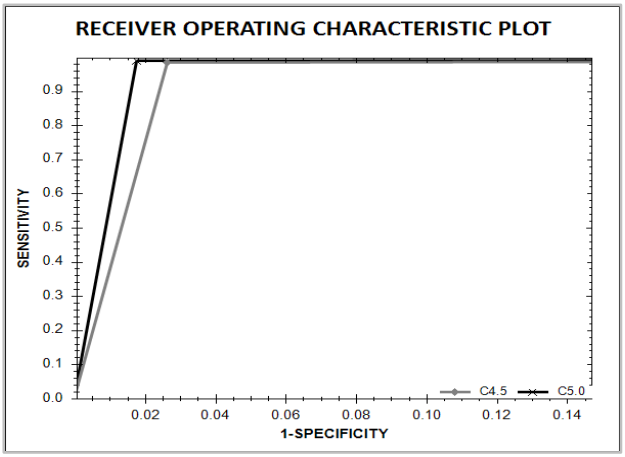

Fig. 1. ROC curve for Data set ID = 1 Training and Testing

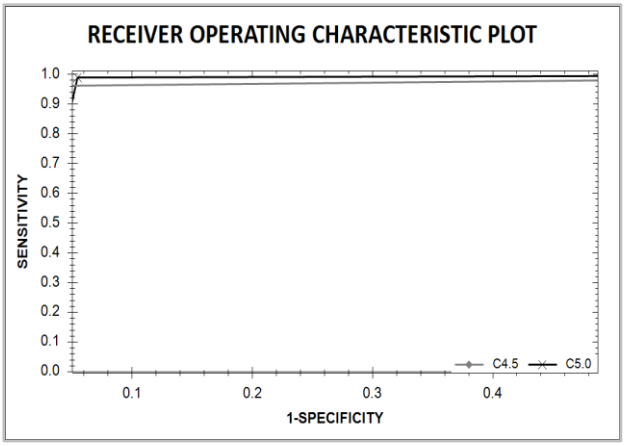

Fig. 2. ROC curve for Data set ID = 2 Training and Testing

Considering the Spam E-Mail Dataset i.e. ($Data\ set\ ID = 2$) the data is vertically portioned and distributed amongst 3 users of the KDLPPDM system. The

training and testing data considered contain 4601 records. The ROC curve for the training and testing results is shown in Fig. 2 of this paper. 133 and 163 records were misclassified when considering the $C5.0$ and $C4.5$ data mining engines. The execution time of $C4.5$ data mining engine based $KDLPPDM$ system was found to be around 4.4 seconds and the execution time of the $C5.0$ based $KDLPPDM$ system was found to be about 0.5 seconds.

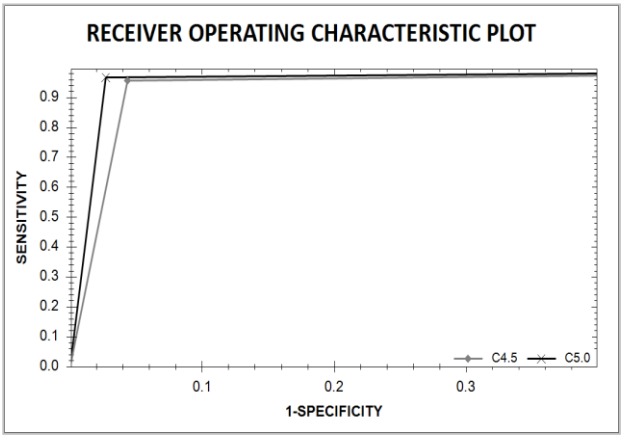

Fig. 3.ROC curve for $Data\ set\ ID = 3$ Training on 300 records

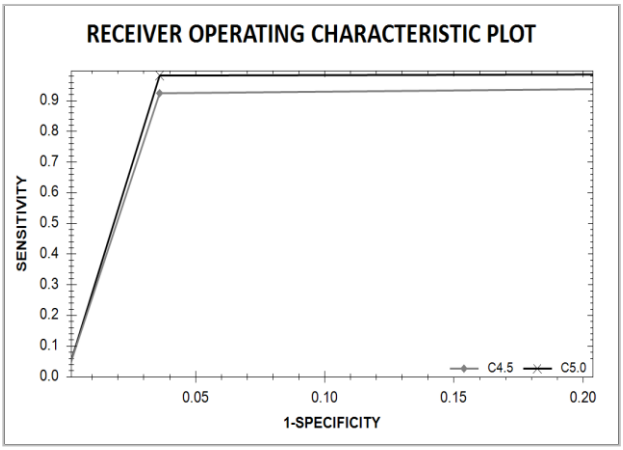

Fig. 4.ROC curve for $Data\ set\ ID = 3$ Testing on 135 records

Considering the $Data\ set\ ID = 3$ the total number of records used for the creation of the set N_{Set} is 300 and the test data set available with the initiator usr_{int} consists of 135 unique records. The classification accuracy represented as ROC curves of the rules within the N_{Set} on the 300 training records is shown in Fig. 3. The rules generated using the $C4.5$ data mining engine based $KDLPPDM$ system was able to classify 287 records accurately. 291 records of the 300 records were accurately classified using the rules generated by the $C5.0$ based $KDLPPDM$ system. On evaluating the N_{Set} rules obtained from both the systems to classify the 135 testing records it was observed that 7 and 3 records were misclassified by the $C4.5$ based system and $C5.0$ based system. The ROC curves obtained is shown in Fig. 4. The area covered by the $C5.0$ ROC curve is 0.96 and the area under the $C4.5$ ROC curve is 0.92.

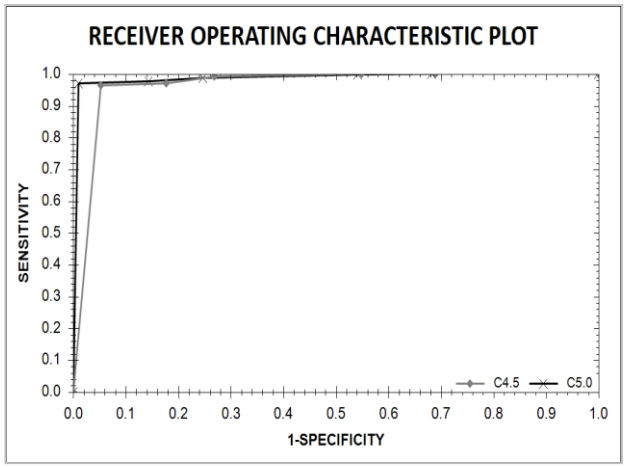

Fig. 5. ROC curve for Data set ID $= 4$ Training on 4435 records

The Landsat Multi-Spectral Scanner Image Data (i.e.$Data\ set\ ID = 4$) consists of a total of 6435 records out of which 4435 records were used for rule generation and 2000 records were used for the creation of the test data set. The accuracy of the classification rules on the train dataset of 4435 records is shown as a ROC plot in Fig.5. The ROC curves obtained for evaluation on the test dataset of 2000 records is shown in Fig.6. The rules generated using the $C4.5$ based $KDLPPDM$ system exhibited a classification accuracy of 93.6% on the training dataset and 85.2% on the test dataset. The rules generated using the $C5.0$ based $KDLPPDM$ system exhibited a classification accuracy of about 96.6% on the training data set and 86.1% on the test data set. The execution time of the $C4.5$ algorithm was found to be 1.16 seconds and of the $C5.0$ algorithm was found to be 0.5 seconds.

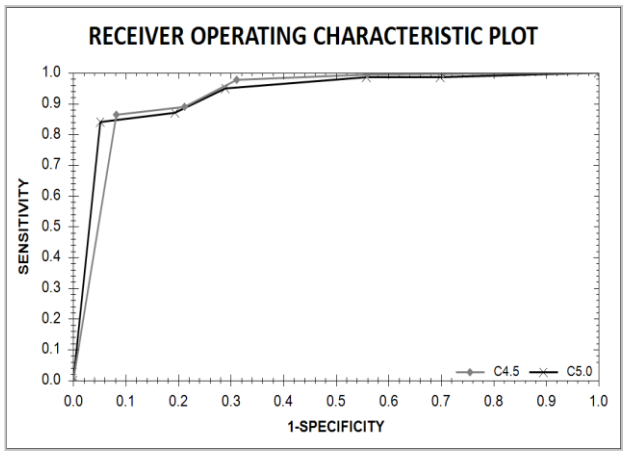

Fig. 6. ROC curve for Data set ID $= 4$ Testing on 2000 records

Of the 690 records in the Credit Approval Data Set (i.e.$Data\ set\ ID = 5$), 490 records were considered for rule generation and the remaining 200 were considered as the

test data set. The $C4.5$ based $KDLPPDM$ system and the $C5.0$ based $KDLPPDM$system were evaluated using this data set. The resultant ROC curves obtained for the training dataset classification and testing dataset classification is shown in Fig. 7 and 8. The execution time of the $C5.0$ based $KDLPPDM$ system was 0.1 seconds when compared to the 0.46 seconds of the $C4.5$ based $KDLPPDM$ system. From the ROC analysis a classification efficiency of 0.88 and 0.80 was achieved by the $C4.5$ based $KDLPPDM$ system on the training and test datasets. A classification efficiency of 0.90 and 0.80 was achieved by the $C5.0$ based $KDLPPDM$ system on the training and test datasets.

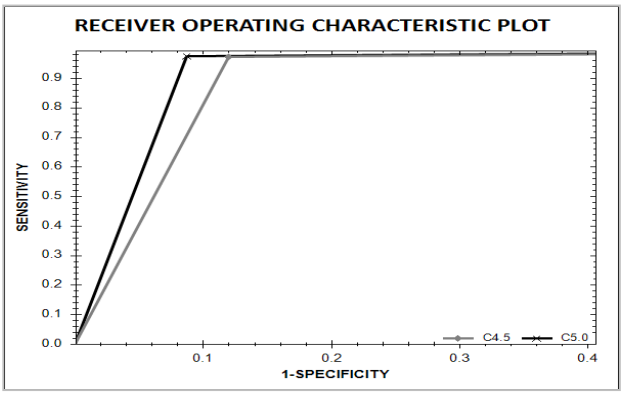

Fig. 7.ROC curve for Data set ID = 5 Training on 490 records

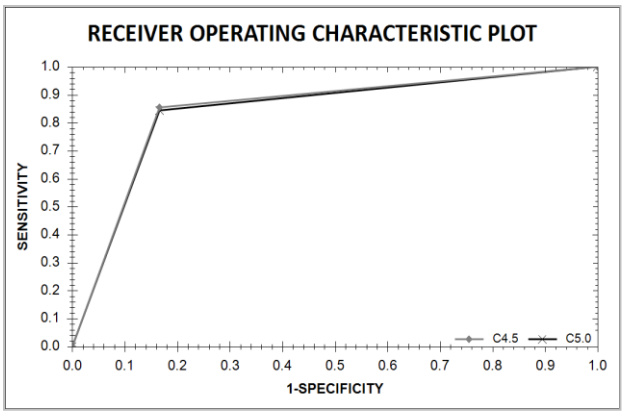

Fig. 8. ROC curve for Data set ID = 5 Testing on 200 records

In the proposed $KDLPPDM$ system no data is published but only the locally generated rules are published to achieve accurate mining results. A large number of classification rules induces and overhead on the system both in terms of communication and computations involved in encryptions and decryptions using the $Commutative\ RSA$ algorithm. The overheads proportional to the number of rules generated.

A study is conducted to study the time taken to generate the classification rules and the size of the combined rule set (i.e. N_{Set}) obtained at each of the 3 user nodes for all the data sets mentioned in Table 6. The results obtained are shown in Fig. 9 and 10. Based on the results obtained it was observed that the $C5.0$ based $KDLPPDM$ is 87.9% efficient in execution when compared to the $C4.5$ based $KDLPPDM$ system.Based on Fig. 10 the number of classification rules generated by the $C5.0$ based $KDLPPDM$ are smaller in number when compared to the number of classification rules generated by the $C4.5$ based $KDLPPDM$ system there by reducing the overheads in communications and computations.

Fig. 9.: Time Taken for Rule Generation considering Varied Data Sets

The proposed KDLPPDM system solely relies on the rules generated to achieve accurate data mining results. The rules generated are evaluated on the datasets discussed above and the results obtained are presented as ROC curves. The greater the area covered by the ROC curves better is the classification accuracy achieved. The area covered by the C4.5 based KDLPPDM system and the C5.0 based KDLPPDM system considering the training and testing data of the 5 datasets is presented in Fig. 11 and 12.

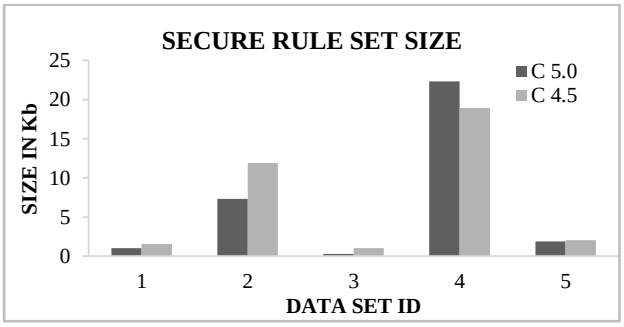

Fig. 10. Secure Rule Set (N_{Set}) Size considering Varied Data Sets

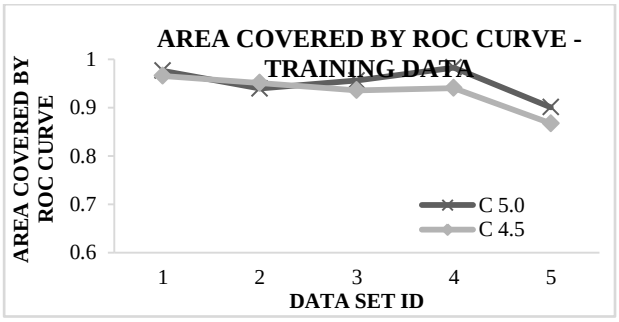

Fig. 11. Area Covered by the ROC curves considering Training Data

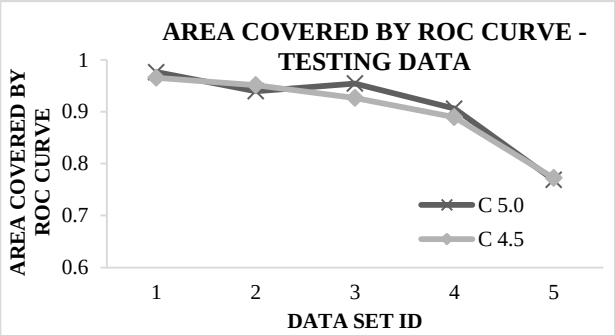

Fig. 12. Area Covered by the ROC curves considering Testing Data

High classification accuracy is a desired feature of any data mining system. To evaluate the classification accuracy of the proposed KDLPPDM system embodying both the C4.5 and the C5.0 algorithm, the five datasets discussed in Table 6 above have been considered. The classification accuracy is evaluated both on the training and testing data of the five datasets. The results obtained are graphically presented in Fig. 13 and 14 of this paper. It is observed that the average classification accuracy of the C5.0 based KDLPPDM system is 96.8% for the train data and 92.5% for the testing data. The average accuracy of the C4.5 based KDLPPDM system is 95.3% for training data and 91.7% for the testing data.

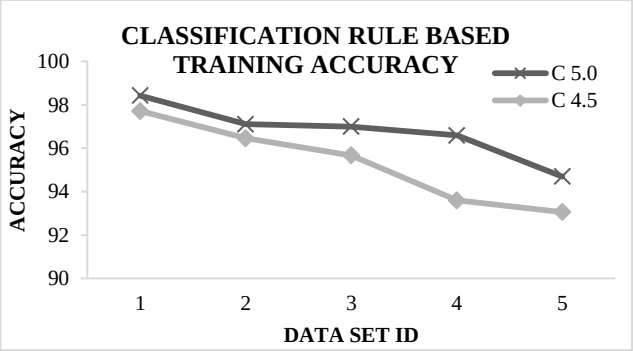

Fig. 13. Classification Accuracy of the KDLPPDM system on the Training Data

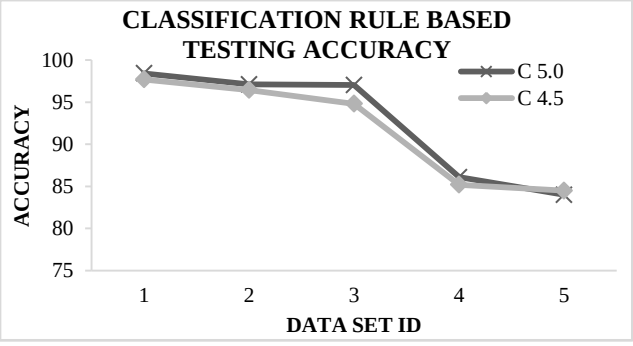

Fig. 14. Classification Accuracy of the KDLPPDM system on the Testing Data

Based on the results presented in this section it can be concluded that the proposed $KDLPPDM$ system achieves privacy of the data and exhibits efficient data mining results on versatile datasets. The performance of the $C5.0$ based $KDLPPDM$ is found to be better in terms of the classification accuracy, overhead reduction and execution times against the $C4.5$ based $KDLPPDM$ system. The results obtain also prove that the $C5.0$ data mining algorithm is robust and efficient when compared to its predecessor the $C4.5$ data mining algorithm hence the authors of the paper consider the adoption of the C5.0 data mining algorithm in the proposed $KDLPPDM$ system.

Conclusion

Large organization and institutions generally have data housed at varied locations. The use of $PPDM$ system is considered in such scenarios to mine the data from distributed resources. The issues that exist in $PPDM$ systems are presented in this paper. The proposed $KDLPPDM$ system is discussed in this paper which overcomes the overheads arising due to key exchange and key computation by adopting the $Commutative\ RSA$ cryptographic algorithm. The $KDLPPDM$ discussed considers no publication of data and propagates the publication of the classification rules generated using the $C5.0$ algorithm by each party. The integrity of rules published towards the construction of the secure rule set are preserved by the $Commutative\ RSA$ algorithm. The $KDLPPDM$ system is established in a three step approach post the initialization step and is discussed in detail. The mining efficiency of the $KDLPPDM$ relies on the classification rules generated by the parties involved. The adoption of the $C5.0$ data mining algorithm in the $KDLPPDM$ system over the $C4.5$ data mining algorithm is justified based on the experimental study presented in this paper.

References

V.S. Verykios *et al*.,(2001),State-of-the-Art in Privacy Preserving Data Mining, *SIGMOD Record*, vol.33, no.1, pp.50-57.

Y. Lindell and B. Pinkas(2000),Privacy preserving data mining, *in Proc. Int'l Cryptology Conference (CRYPTO)*, pp. 36-54.

Yaping Li, Minghua Chen, Qiwei Li and Wei Zhang , Enabling Multi-level Trust in Privacy Preserving Data Mining(2012), *in IEEE Transactions on Knowledge and Data Engineering* ,vol.24,no.09,pp.1598 – 1612.

O. Goldreich, "Secure multi-party computation," *Final (incomplete) draft, version 1.4*, 2002.

A.W.-C. Fu, R. C.-W.Wong, and K.Wang,(27-30 Nov. 2005),Privacy-preserving frequent pattern mining across private databases, *Fifth IEEE International Conference on Data Mining*.

D. Agrawal and C. C. Aggarwal,(2001),On the design and quantification of privacy preserving data mining algorithms, *in Proceeding of the 20th ACM Symposium on Principles of Database Systems, Santa Barbara, California*, pp.247-255.

K. Chen and L. Liu,(27-30 Nov. 2005),Privacy preserving data classification with rotation perturbation, *Fifth IEEE International Conference on Data Mining*.

S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu(2007),Time series compressibility and privacy, *in Proc. Int'l Conf. on Very Large Data Bases*, pp. 459-470.

L. Sweeney, "k-anonymity: A model for protecting privacy,(2002),*International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS)*, vol. 10,no. 5,pp. 557-570.

C. C. Aggarwal and P. S. Yu,(2004),A condensation approach to privacy preserving data mining, *in Proc. Int'l Conf. on Extending Database Technology (EDBT)*, vol. 2992, pp. 183-199.

Slava Kisilevich, Lior Rokach, Yuval Elovici and Bracha Shapira,(MARCH 2010),Efficient Multidimensional Suppression for K-Anonymity, *IEEE Transactions on Knowledge and Data Engineering* ,vol.22, no.3,pp.334 – 347.

W. Du and Z. Zhan,(2003),Using randomized response techniques for privacy-preserving data mining, *in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.505-510.

R. Agrawal, R. Srikant, and D. Thomas,(June,2005),Privacy preserving OLAP, *in Proc. ACM SIGMOD Int'l Conf. on Management of Data*. http://www.rulequest.com/see5-info.html

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu and Philip S. Yu, *et al*.,(2008), Top 10 algorithms in data mining, Springer, *Knowledge and Information Systems*, vol.14, no.1 pp.1-37.

Tomasz Bujlow, Tahir Riaz, Jens Myrup Pedersen,(Jan. 30 2012-Feb. 2 2012),A method for classification of network traffic based on C5.0 Machine Learning Algorithm, *in Workshop on Computing, Networking and Communications, IEEE*, pp. 237-241.

Po-Hsun. Sung, Jyh-Dong Lin, Shih-Huang Chen, Shun-Hsing Chen, and Jr-Hung Peng,(7-10 Dec. 2010), Utilization of Data Mining on Asset Management of Freeway Flexible Pavement, *Proceedings of the Industrial Engineering and Engineering Management (IEEM - IEEE)*,pp.977 – 979.

Ming-Jun Xiao, Kai Han, Liu-Sheng Huang and Jing-Yuan Li,(Oct. 2006) Privacy Preserving C4.5 Algorithm Over Horizontally Partitioned Data, *Proceedings of the Fifth IEEE International Conference on Grid and Cooperative Computing(GCC 2006)*, pp.78 – 85.

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, (2006),Our data, ourselves: Privacy via distributed noise generation, *Advances in Cryptology-EUROCRYPT 2006*, pp. 486–503.

M. Islam , L. Brankovic,(2003),Noise Addition for Protecting Privacy,*in Data Mining,Proceedings of The 6th Engineering Mathematics and Applications Conference,(EMAC2003), Sydney*, pp. 85–90.

Matthews, Gregory J., Harel, Ofer,(2011),Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy, *Statistics Surveys*, vol. 5, pp. 1-29.

J. Vaidya and C. W. Clifton,(2002),Privacy preserving association rule mining in vertically partitioned data, *in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*,pp. 639-644.

R. Agrawal and R. Srikant,(2000),Privacy preserving data mining, *in Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pp. 439-450.

B. Pinkas,(2002), Cryptographic techniques for privacy-preserving data mining, *ACM SIGKDD Explorations Newsletter*, vol.4,no.2, pp. 12-19.

Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu(December 2002),Tools for Privacy Preserving Distributed Data Mining,*SIGKDD Explorations*, vol. 4, no. 2, pp. 28-34.

Alexandre Evfimievski and Tyrone Grandison ,(2009),Privacy-Preserving Data Mining, *IGI Global*, pp. 1-8.

Victor P. Hubenko Jr., Richard A. Raines, Rusty O. Baldwin, Barry E. Mullins, Robert F. Mills, and Michael R. Grimaila,(July-August 2007),Improving Satellite Multicast Security Scalability by Reducing Rekeying Requirements, *IEEE Network*,vol.21,no.4, pp.51-56.

Bezawada Bruhadeshwar and Sandeep S. Kulkarni,"Balancing Revocation and Storage Trade-Offs in Secure Group Communication,(Jan.-Feb. 2011), , *IEEE Transactions on Dependable and Secure Computing* , vol. 8,no. 1, pp.58-73.

Nathaniel Karst, and Stephen B. Wicker,(Oct. 2012)"On the Rekeying Load in Group Key Distributions Using Cover-Free Families, *IEEE Transactions on Information Theory*,vol.58, no.10,pp.6667 – 6671.
Piotr Andruszkiewicz, Marzena

Kryszkiewicz(2011),Privacy Preserving Classification and Association Rules Mining over Centralised Data, *Warsaw*.

Patrick Sharkey, Hongwei Tian, Weining Zhang, and Shouhuai Xu,(2008),Privacy-Preserving Data Mining through Knowledge Model Sharing, *Springer-Verlag Berlin Heidelberg ,Lecture Notes in Computer Science(LNCS)*,vol.4890, pp. 97-115.

Pui K. Fong and Jens H. Weber-Jahnke,(2012),Privacy Preserving Decision Tree Learning Using Unrealized Data Sets, *IEEE Transactions on Knowledge and Data Engineering*,vol.24,no.2, pp.353 – 364.

Kumaraswamy S is currently working as an Assistant Professor in the Department of Computer Science and Engineering, KNS Institute of Technology, Bangalore, India. He obtained his Bachelor of Engineering from SiddaGanga Institute of Technology, Bangalore University, Tumkur. He received his M E Degree in Computer Science and Engineering from UVCE, Bangalore University, Bangalore. He is a research scholar in Bangalore University. He has 14 years of teaching experience His research interest is in the area of Data mining, Web mining , Semantic web and cloud computing.

S H Manjula is currently Associate Professor, Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. She obtained her Bachelor of Engineering, Masters of Engineering and Ph.D in Computer Science and Engineering. She published book on Wireless sensor Networks. She published more than 30 papers. She refereed international journal and conference papers. Her research interests are in the field of Wireless Sensor Networks, Semantic web and Data mining.

Venugopal K R is currently Special Officer, DVG Bangalore University and Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D. in Economics from Bangalore University and Ph.D.in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited 39 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc.. During his three decades of service at UVCE he has over 400 research papers to his credit. He was a Post Doctoral Research Scholar at University of Southern California, USA. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and DistributedSystems, Digital Signal Processing and Data Mining.

L M Patnaik is a Ex-Vice Chancellor, Defense Institute of Advanced Technology, Pune, India. He was a Professor since 1986 with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 700 research publications in refereed International Journals and refereed International Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. Currently he is Honorary Professor, Indian Institute of Science Bangalore, India. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.