Classification Technique with Privacy Preservation of Outsourced Database

Ishita Sharma 1, Mr.Shrikant Lade 2
1 (M.Tech-IV Sem) RKDF IST, Bhopal,
2 ( HOD CS& IT) RKDF IST, Bhopal

Abstract— With the easy availability of internet facility number of its user are increasing day by day. Most of the servers provide different type of services to users. But this leads to one problem which is related to user data security. So keeping this goal in mind server will get perturbed data then perform classification on that perturbed data. Once classification done then result send back to the user. Security is maintained at sender and receiver side, by using cryptosystem, where textual data is send to the server for classification. Results are compare with previous approach on different parameters, which shows that proposed work is efficient.

Keywords:-Classification, Cryptosystem, Privacy Preserving Mining.

I. INTRODUCTION

Data mining is a technique that deals with the extraction of hidden predictive information from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking relevant information. Data mining extract knowledge using tools which predict future trends and methods for decision-making, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data doubling each year more data is gathered and data mining is becoming an increasingly important tool to transform this data into information. Long process of research and product development evolved data mining in current scenarios. This evolution began when business data was first stored on computers, data access and prediction, business policy and more recently, generated technologies that allow users to navigate through their data in real time.

One of the fields of data mining is Privacy-Preserving Data Mining (PPDM) key goal in most distributed methods for privacy preserving data mining (PPDM) is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. The participants/individual may wish to collaborate in obtaining aggregate results, but may not have full trust on each other in terms of the sharing of their own data sets.

Recent advances on server computing replaces many traditional techniques and provides various services to the clients over the Internet in a flexible manner (i.e., ondemand, pay-per use) [1]. This leads to a new paradigm of service where a server in a cloud could offer data classification to clients. In particular, the server can automatically process and classify the clients’ data samples remotely based on privately owned training data samples.

The importance of cloud computing in the data classification can be categorized as follows:

1. the cloud is responsible for maintaining and updating training data set for classification.
2. The cloud provides data classification as a service to any clients via Internet while preserving the privacy of clients’ data
3. The cloud helps to offload substantial amount of computation of clients
Classification Technique with Privacy Preservation of Outsourced Database

Paper Outline: This section has given an overview of the motivation and scope of this research purpose. Second section provides a brief literature survey. Third section presents the proposed work with different related terms and steps. Fifth section discussed the evaluation matrices and experimental results analysis. Sixth section concludes the paper by presenting a conclusion and suggest direction for future research.

II. RELATED WORK

Privacy preserving data mining has become a hot spot in data mining research. The main reasons behind it are the importance of private data, enhanced technology allowing ease of storage, access, transfer, manipulation of centralized and distributed data. To save it from unauthorized access and attacks to get the knowledge many perturbation techniques have been used by various researchers. The attacker can have basic information regarding the dataset. The important distinction between our scenario and others is that, in ours, the results as well as sensitive attributes are not intended to be open to others like that in [3]. There are many techniques that are prevalent for privacy preserving data mining. The literature in Majority of the work in the literature were developed for the distributed setting where different parties hold parts of the training data sets and securely train a common classifier without each party needing to disclose its own training data to other parties [5].

After the training, each party holds part of the classification parameters. In order to classify a new test sample, each party has to be involved equally to compute part of the kernel matrix and then all parties together, or the trusted thirdparty will classify the test sample.

The works in [6] exploited the secure multi-party integer summation in order to compute the kernel matrix. Basically, each party generates a Gramm matrix using scalar products of training and test data samples. This Gramm matrix is later revealed to the trusted third party who will compute the kernel matrix and then classify the test sample. Revealing the Gramm matrix may leak the private data and, therefore, privacy cannot be entirely preserved.

The work in [4] proposed for the first time a strongly privacy-enhanced protocol for SVM using cryptographic primitives where the authors assumed that the training data is distributed. Hence, in order to preserve the privacy they developed a protocol to perform secure kernel sharing, prediction and training using secret sharing and homomorphic encryption techniques. At the end of the training, each party will hold a share of the secret.

In [16] a client sends the input data sample in an encrypted format to the server. Then the server exploits the homomorphic encryption properties to perform the operations directly on the encrypted data sample. If there are any operations that cannot be handled by the homomorphic properties, then there will be a limited amount of interaction between the client and server based on two-party secure computation protocol. This work assume that both the client and the server will execute the protocol correctly in order to maintain their reputation, hence they will behave in a semi-honest manner, i.e., they are honest but curious, so privacy is a real issue.

III. PROPOSED WORK

As the privacy of dataset is important for storing it at different stations for ease of access, which is done in variety of ways.

In order to put this dataset on the server for different purpose it needs protection from unauthorized user who uses it for unfamiliar activities.

For this method need for perturbing the dataset is proposed in this work. Process of perturbation start from the pre-processing of the dataset which make dataset in the required format for the working of the environment.
Whole work is divide into two part first is at server side while other is at user side. First when user want to classify its data it will follow below steps:

**Pre-Processing**: Text preprocessing is consisting of words which are responsible for lowering the performance of learning models. Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop word elimination and stemming. Stop-words are functional words which occur frequently in the language of the text (for example a, the, an, of etc. in English language), so that they are not useful for classification. Here we read whole project and put all words in the vector. Now again read the file which contains stop words then remove similar words from the vector. Once the data is pre-process then it will be the collection of the words that may be in the vector. For example let one document is taken and its text vector is $Rd[] = \{a1, f1, s1, a2, s2, a3, a4, f2, \ldots \ldots \ldots \ldots \ \}$ and let the stop words collection is $S[] = \{a1, a2, a3, \ldots \ldots \ \}$ . Then the vector obtain after the Pre-Processing is $D[] = \{f1, s1, s2, f2, \ldots \ldots \ldots \}$.

$$D[] = Rd[] – D[]$$

For Example: $Rd[] = \{\text{‘Every’}, \text{‘morning’}, \text{‘Ram’}, \text{‘study’}, \text{‘for’}, \text{‘two’}, \text{‘hour’}, \text{‘and’}, \text{‘during’}, \text{‘this’}, \text{‘time’}, \text{‘his’}, \text{‘mother’}, \text{‘give’}, \text{‘him’}, \text{‘one’}, \text{‘glass’}, \text{‘milk’}, \text{‘with’}, \text{‘bread’}, \text{‘jam’}, \text{‘in’ \ ‘breakfast’} \}$

After pre-processing

Now $D[] = \{\text{‘Ram’}, \text{‘hour’}, \text{‘time’}, \text{‘glass’}, \text{‘milk’}, \text{‘bread’}, \text{‘jam’}, \text{‘breakfast’} \}$

Now assign number to each text of the different document. So that a dictionary of words with there number is created where each text is identified by separate number. Such as

$D[] = \{1, 2, 3, 4, 6, 7, 8, 9\}$

So for n document has its own vector sequence $D[n]$. 

**Pailler Encryption**: This cryptosystem is base on the public and private key concept. Here input vector $D[n]$, will be encrypt by this algorithm.

1. Choose two large prime numbers $p$ and $q$ randomly and independently of each other such that $\gcd(pq,(p-1)(q-1)) = 1$.

2. Compute RSA modulus $n = pq$ and
Classification Technique with Privacy Preservation of Outsourced Database

Carmichael’s function \( \lambda = \text{lcm}(p-1, q-1) \)

3. Select generator \( g \). Select \( \alpha \) and \( \beta \) randomly from a set \( Z^*n \) then calculate \( g = (\alpha \beta) \mod (n^*n) \)

4. Calculate the following modular multiplicative inverse

\[
\mu = \text{mod}(n) \div (Lg \lambda \mod(n^*n) - 1)
\]

Where the function \( L \) is defined as \((u) = (u - 1)/n\).

So The public key is \((n, g)\), private key is \((\lambda, \mu)\).

**Normalization:**

This is done at server end, where server is not able to understand the data as it is in encrypted form. This normalization step is required as numbers need to convert into same platform if it is in different level.

![Fig.2 Block diagram of proposed work at server side.](image)

\[
X = (X_i - X') / (\sigma \cdot \sigma)
\]

where \( X, X' \) and \( \sigma \) denote the individual value, mean and standard deviation.

**Decision value:**

The encrypted test sample \( \frac{1}{2}gt \) is used to compute the polynomial kernel \( K_p = [(\gamma X_i)'x(T) + (\gamma^*Y)] \) in the ED. Its power is raise by \( p \) for the polynomial equation. In [base] this is done by client where it send the value from server \( t \) client then client raise its power and send it back. So for exchanging this information one has to encrypt data then send and decrypt for performing other operations. This step is remove in proposed work so that server time will be safe.

Finally sum all the values for generating the decision value of the work.

\[
d(t) = \sum K_p
\]

**Classification:**
As the decision function generates a value which is termed as decision value with a sign that will help in classifying the data, here the base on the positive or negative value of the decision value. Document is classified into two classes.

IV. EXPERIMENT AND RESULT

This section presents the experimental evaluation of the proposed. All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on an 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional. Experiment done on the Image dataset which have collection of different images and textual data.

Evaluation Parameter

Accuracy

In this parameter the number of classification done by system correctly is divide by total number of correct and incorrect classification.

\[
\text{Accuracy} = \frac{\text{Correct}}{\text{Correct} + \text{Incorrect}} \times 100
\]

Execution time

As the work done on the important resource that is server so execution time should be less as possible. So this is a very important parameter to evaluate this work.

Results

Space is same for both the propose work as well as previous work as both are following similar steps at the sender end. Perturbation done in the original dataset before sending to the server.

<table>
<thead>
<tr>
<th>Sets</th>
<th>Accuracy</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77.7778</td>
<td>22.843</td>
</tr>
<tr>
<td>2</td>
<td>84.1270</td>
<td>21.877</td>
</tr>
<tr>
<td>3</td>
<td>84.1270</td>
<td>19.107</td>
</tr>
<tr>
<td>4</td>
<td>74.6032</td>
<td>23.9925</td>
</tr>
<tr>
<td>5</td>
<td>74.6032</td>
<td>27.4377</td>
</tr>
</tbody>
</table>

Table 2. Accuracy result on From the previous work.
Classification Technique with Privacy Preservation of Outsourced Database

<table>
<thead>
<tr>
<th>Sets</th>
<th>Modify</th>
<th>Accuracy</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>77.7778</td>
<td>4.8128</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>84.1270</td>
<td>4.8486</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>84.1270</td>
<td>5.4436</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>74.6032</td>
<td>4.6714</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>74.6032</td>
<td>4.3684</td>
</tr>
</tbody>
</table>

Table 3. Accuracy result on From the proposed work.

From above table we can observe that the combination of different features are classify in one to all fashion the number of class are different for the insert images so the accuracy also varies. But it has been observed that both proposed work and previous work accuracy is same although some of steps are done at server end in previous work. Due to this modification execution time get sharply decrease.

<table>
<thead>
<tr>
<th>Sets</th>
<th>Modify</th>
<th>Accuracy</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>77.77</td>
<td>6.9031</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>55.5556</td>
<td>4.0167</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>55.5556</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Table 2. Accuracy result on the proposed work.

From above table we can observe that the combination of different features are classify in one to all fashion the number of class are different for the insert images so the accuracy also varies. But it has been observed that both proposed work and previous work accuracy is same although some of steps are done at server end in previous work. Due to this modification execution time get sharply decrease.

From above graph it is also observed that the execution time in the proposed algorithm is less as compared to the previous work[16]. So work done for privacy preserving is good in all sense as compared to the previous work done in [16] in all aspects.

V. Conclusion

As different services are launched by the servers, but there may chance of data manipulation or delicacy. This paper provide privacy for classification data services. Here combination of classification and encryption technique leads to develop one robust approach. It has been observed from the results that proposed work will successfully classify both images as well as textual data. There is always work remain after every steps, so number of services are still remaining for which security is required.
References


[16] Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud Yogachandran Rahulamathavan, Member, IEEE, Raphael C.-W. Phan, Suresh Veluru, Kanapathippillai Cumanan, Member, IEEE, and Muttukrishnan Rajarajan, Senior Member, IEEE. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 11, NO. 5, SEPTEMBER/OCTOBER 2014