



A Robust Technique for Privacy Preservation of Outsourced Database

¹Umesh Kumar, ²Diamond Jonawal

¹M.Tech (CTA) Research Scholar

^{1,2}RGTU University, Bhopal, India

Abstract— As the number of internet users are increasing day by day, so the availability of different servers to serve users for their need is come in demand. This rises to common problem of data privacy against various data miners. As data mining will lead to fetch information from the raw data, so this paper presents a robust technique for the miner to provide privacy of the dataset. Security is maintained at sender and receiver side by using Paillier cryptosystem, where textual data is send to the server for classification. Results are compared with previous approach on different parameters, which shows that proposed work is efficient.

Keywords— Privacy Preserving Mining, Association Rule Mining, Data Perturbation, Paillier cryptosystem, Web Usage 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). The key goal in most distributed methods of 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.

A recent advance on server computing replaces many traditional techniques and provides various services to the clients over the Internet in a flexible manner (i.e., on demand, 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

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 discusses the evaluation matrices and experimental results analysis. Sixth section concludes the paper by presenting a conclusion and suggests 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 third-party 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 assumes 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.

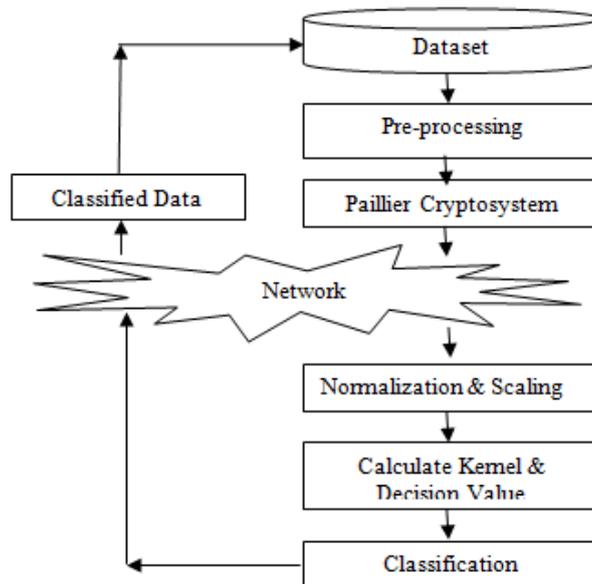


Fig 1: Block Diagram of Proposed Work

Whole work is divide into two parts- first is at server side while other is at user side. First when user wants to classify its data it will follow the below steps:

A. Pre-Processing:

Text pre-processing is consisting of words which are responsible for lowering the performance of learning models. Data pre-processing 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 $Rd[] = \{a1, f1, s1, a2, s2, a3, a4, f2, \dots, an\}$ and let the stop words collection is $S[] = \{a1, a2, a3, \dots, am\}$. Then the vector obtain after the Pre-Processing is $D[] = \{f1, s1, s2, f2, \dots, fx\}$.

$$D[] = Rd[] - S[]$$

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

After pre-processing

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

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 $D[n]$.

B. Paillier Cryptosystem:

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 Carmichael's function $\lambda = \text{lcm}(p-1, q-1)$
3. Select generator g , Select α and β randomly from a set $\mathbb{Z}_n^* \times \mathbb{Z}_n^*$ then calculate $g = (\alpha n + 1) \beta^* \beta \text{ mod } (n^* n)$
4. Calculate the following modular multiplicative inverse
 $\mu = \text{mod}(n) / (L(g \lambda \text{ mod } (n^* n))^{-1})$
Where the function L is defined as $L(u) = (u-1)/n$.

So the public key is (n, g) , private key is (λ, μ) .

C. Normalization & Scaling:

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 requiring as numbers need to convert into same platform if it is in different level.

$$X = (X_i - X') / (\sigma * \sigma)$$

Where X , X' and σ denote the individual value, mean and standard deviation.

While in scaling the normalized value is multiply by a constant.

$$T \Rightarrow YX = Y[(X_i - X') / (\sigma * \sigma)]$$

D. Kernel & Decision value:

The encrypted test sample $\frac{1}{2}gt$ is used to compute the polynomial kernel $K_p = [(YX_i)^x(T) + (Y^*Y)]$ 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 removed in proposed work so that server time will be safe.

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

$$d(t) = \sum K_p$$

E. Classification:

As the decision function generate a value which is term as decision value has sign which 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.

```
If  $d(t) > 0$ 
    +class
Else
    -class
end
```

IV. EXPERIMENT AND RESULT

This section presents the experimental evaluation of the proposed perturbation and de-perturbation technique for privacy prevention. To obtain AR this work used the Apriori algorithm [1], which is a common algorithm to extract frequent rules. All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on a 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional. Experiment done on the customer shopping dataset which have collection of items, cost, total amount, etc. attributes.

Dataset:

In order to evaluate the proposed work, two different type of dataset is use for the classification and privacy analysis. First database JAFFE contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. Second artificial dataset is for text classification where different class of document are classify then separate into their class for experiment.

Evaluation Parameter:

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

$$\text{Accuracy} = [\text{Correct} / (\text{Correct} + \text{Incorrect})] * 100$$

2. **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.
3. **Dataset Size:** Here size of dataset is analysed while after perturbing. As if the size increases then it require more space to store it on the server.

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 1. Comparison with Minsup 0.2 (with MSNBC dataset)

Accuracy of JAFFE Dataset				
	Previous Paper		Modify	
Feature	Accuracy	Time	Accuracy	Time
1	65.0794	47.483	65.0794	7.168
2	68.254	33.607	68.254	8.114
3	65.0794	21.107	65.0794	7.094
4	49.2063	21.247	49.2063	8.107
5	52.381	32.997	52.381	6.513
6	61.9048	31.028	61.9048	7.418

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 2. Accuracy results from the proposed work.

Accuracy of Text Dataset		
	Modify	
Feature	Accuracy	Time
1	50	0.0329
2	16.6667	0.0167
3	66.6667	0.014
4	50	21.2465

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.

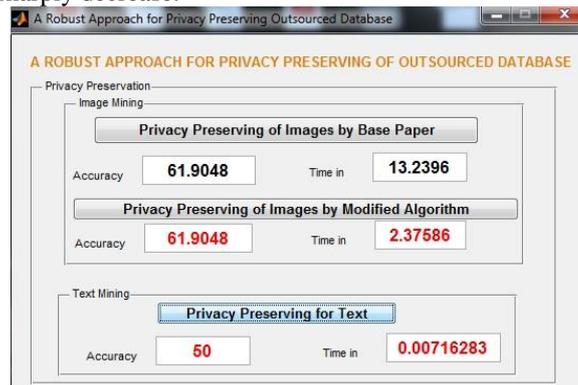


Fig 2: GUI in MatLab

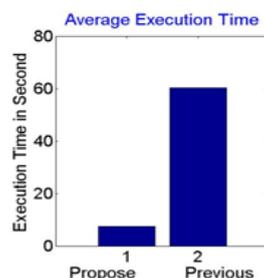


Fig 3: Performance Graph

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 AND FUTURE WORK

Preserving privacy in data mining activities is a very important issue in many applications. Randomization-based techniques are likely to play an important role in this domain. In this paper, a new approach to solve the problem of privacy preserving data mining in the scenario of outsourced business transaction database. This approach is efficient and better than previous work which is shown in results. A new approach of text classification base on the cryptosystem is developed. Accuracy results both in image classification and text document classification is efficient. In future, work need to make it more powerful for distributed databases as will.

REFERENCES

- [1] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects" in IEEE 2012 Third International Conference on Computer and Communication Technology.
- [2] R. Agrawal, T. Imielinski, and A. N. Swami. "Mining Association Rules between Sets of Items in Large Databases". SIGMOD 1993.
- [3] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases" in IEEE SYSTEMS JOURNAL, VOL. 7, NO. 3, SEPTEMBER 2013.
- [4] H. Lipmaa, S. Laur, and T. Mielikainen, "Cryptographically Private Support Vector Machines," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 618-624, Aug. 2006.
- [5] H. Yu, X. Jiang, and J. Vaidya, "Privacy-Preserving SVM using Nonlinear Kernels on Horizontally Partitioned Data," Proc. ACM Symp. Applied Computing (SAC), 2006.
- [6] H. Yu, J. Vaidya, and X. Jiang, "Privacy-Preserving SVM Classification on Vertically Partitioned Data," Proc. 10th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), 2006.
- [7] K. Chen and L. Liu, "Privacy Preserving Data Classification with Rotation Perturbation," Proc. IEEE CS Fifth Int'l Conf. Data Mining (ICDM), pp. 589-592, 2005.
- [8] K.-P. Lin, M.-S. Chen, "On The Design and Analysis of the Privacy-Preserving SVM Classifier," IEEE Trans. Knowledge and Data Engineering, vol. 23, no. 11, pp. 1704-1717, Nov. 2011.
- [9] Zhengyou Zhou, Liusheng Huang, Ye Yun, "Privacy Preserving Attribute Reduction Based on Rough Set", in IEEE, 2009, Second International Workshop on Knowledge Discovery and Data Mining.
- [10] Xiaolin Zhang, Hongjing Bi, "Research on Privacy Preserving Classification Data Mining Based on Random Perturbation", in, International Conference on Information, Networking and Automation (ICINA), 2010
- [11] F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-preserving outsourcing of association rule mining," ISTI-CNR, Pisa, Italy, Tech Rep. 2009-TR-013, 2009.
- [12] Nikunj H. Domadiya, Udai Pratap Rao, "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database", in IEEE Journal, 2012.
- [13] Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, "Privacy Preserving Decision Tree Mining from Perturbed Data", in, Proc. of 42nd Hawaii International Conference on System Sciences – 2009.
- [14] Fang Lu, Wei-jun Zhong, Yu-lin Zhang, Shu-e Mei, "Privacy-preserving Association Rules Mining Using the Grouping Unrelated-question Model", in IEEE Journal, 2007.
- [15] R. Mahesh, T. Meyyappan "Anonymization Technique through Record Elimination to Preserve Privacy of Published Data", in Proc. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, Feb. 21-22.
- [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
- [17] Yingjie Wu, Shangbin Liao, Xiaowen Ruan, Xiaodong Wang, "Privacy Preservation in Transaction Databases based on Anatomy technique", in IEEE International Conference on Computer Science & Education, 2010