International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 02, Issue: 04 (JULY-AUGUST, 2016)

A Robust Algorithm for Classification of Multiple Types of Data

Manisha Kannavdiya[1], Prof. Kailash Patidar[2], Prof. Manoj Yadav[3]
Department of Computer Science, SSSIST, Sehore, India
[1]manisha.kannavdiya@gmail.com, [2]Kailashpaditar123@gmail.com,[3]manoj5283@gmail.com

Abstract — with the increase in different internet services number of users are also increasing. While taking service user may be on risk for sharing data. So this work focuses on increasing the security of the user data while taking classification service. Here algorithm provide robustness by encrypting the data and send to server, while server classify the data in encrypted form. One more security issue is that instead of transferring whole encrypted data, features are extract from the data first then encrypt and send to server for classification. Here proposed work successfully classifies all type of user data in form of text, image, and numeric.

Keyword — Classification, Feature extraction, Encryption.

I. INTRODUCTION

In recent years a new term has evolved call "Cloud" which is provided by different provides, and which is nothing but facility or service of different resources or apparatus like platform, hardware, software, storage's etc., and this make user free from maintenance which has increase the importance of the work as all these are the cloud service provider responsibility. Now to provide such service to the client, naturally the provider's must have and rather can have access to resources which are used by the people/clients. Among the reasons these access are greatly required are for maintenance perspective. As thousands of client are using those service, so infrastructure tends to be capable for making support of this work. In cloud 24x7 Service availability, data maintenance between various devices, then availability of data via any devices, web browser based connectivity. Now since the info gets shared or stored in providers area, the client gets worried about privacy of its data, although there are certain agreements and SLA which are agreed by cloud provider and client. In normal condition client can share data which need to be secure and less expensive. This work will focus on data where user information details of any business/company/organization are considered to be very sensitive and must be confidential. Therefore if the little scales company thinks of using the services like classification. Classifying all account/finance related information on server makes it prone to leakage of sensitive information tells un-authorized users. Therefore securing this finance data is vital before it gets uploaded to the storage and just in case the data stored in server storage gets tampered there should be a method to verify the integrity of the data, moving further specific band of people should have access to this data which may be folks from finance department of client company or special auditors. Simply speaking the client must have the ability to store the data securely, verify the integrity of the data, and share the data securely with specific band of people.

II. RELATED WORK

In [1] Dan Boneh proposed a short group signature algorithm with length below 200 Bytes although the security of the signature is almost same as the standard RSA algorithm. For providing the Group signature privacy proposed scheme utilize the Strong Diffie-Hellman (SDH) hypothesis and a new hypothesis in bilinear groups called the Decision Linear assumption. This scheme stands on a novel Zero Knowledge Proof of Knowledge (ZKPK) of the answer to an SDH trouble where ZKPK is transformed to a cluster signature via the Fiat-Shamir heuristic. K-Nearest Neighbor (K-NN) K-Nearest Neighbor (K-NN) classifier is one of the simplest classifier that discovers the unidentified data point using the previously known data points (nearest neighbor) and classified data points according to the voting system [7]. Consider there are various objects. It would be beneficial for us if we know the characteristics features of one of the objects in order to predict it for its nearest neighbors because nearest neighbor objects have similar characteristics. The majority votes of K-NN can play a very important role in order to classify any new instance, where k is any positive integer (small number). It is one of the most simple data mining techniques. It is mainly known as Memory-based classification because at run time training examples must always be in memory. Euclidean distance is calculated when we take the difference between the attributes in case of continuous attributes. But it suffers from a very serious problem when large values bear down the smaller ones. Continuous attributes must be normalized in order to take over this major problem. In [3] Nahar et al., used predictive apriori approach for generating the rules for heart disease patients. In this research work rules were produced for healthy and sick people. Based on these rules, this research discovered the factors which caused heart problem in men and women. After analyzing the rules authors conclude that women have less possibility of having coronary heart disease as compare to men. In [4] Shouman et al., used K-NN classifier for analyzing the patients suffering from heart disease. The data was collected from UCI and experiment was performed using without voting or with voting K-NN classifier and it was found that K-NN achieved better accuracy without voting in diagnosis of heart diseases as compared to with voting K-NN. In [2] Abdi et al., was constructed a PSO based SVM model for identifying erythema to-squamous diseases which consists two stages. In the first stage optimal feature were extracted using association rule and in second phase the PSO was used to discovered best kernel parameters for SVM in order to improve the accuracy of classifier model. In [5] Schulam et al., proposed a Probabilistic Subtyping Model (PSM) which was mainly designed in order to discovered subtypes of complex, systematic diseases using longitudinal clinical markers collected in electronic health record (EHR) databases and patient registries. Proposed model was a model for clustering time series of clinical markers obtained from routine visits in order to identify homogeneous patient subgroups.

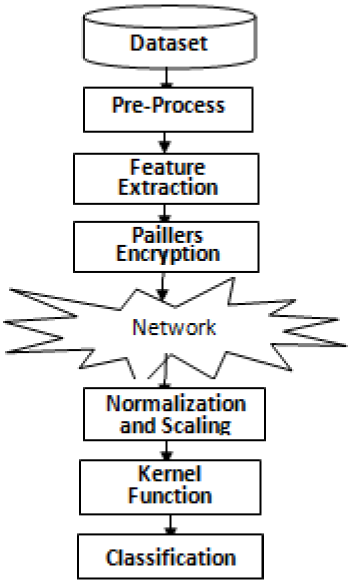

Figure 1 Proposed work block diagram

III. PROPOSED WORK

In this proposed work classification of different type of user data is done at server side while data privacy is done at client side. Although in order to increase the security of the data features of the data are extract and then send those to the server. Whole work is shown in figure 1 block diagram. Dataset: This is collection of unclassified data at the client side which is raw. So it needs to be processed first for applying the security of the data. As classification was done on all type of data so feature extraction and pre-processing of the data is done in their respective way.

3.1 Pre-Processing

Three types of data can be classified in this work first is text, numeric, image.

3.1.1 Text Document

So in case of text data filtration of some stop words are done in this work. This can be understand as let text document contain a sentence S= {I live in a great country of whole world}, then stop words in that sentence are {I, in, a, of}. So removing of those words from the sentence is term as Stop word removal which is a pre-processing technique of text mining. Here collection of these words in a single vector was done. Each document has its own vector which is term as Bag of Words.

3.2 Feature Extraction

In this step each Bag of words contains some of repeated words so Term frequency is calculated for each word of the document. Now those terms which cross minimum frequency threshold act as feature dimension for the document? Now assign unique number to all words which cross that threshold. So each word is identified by that unique number set.

3.3 Numeric Data

In Numeric data pre-processing is done by converting the values in string form to double type. As dataset contain value in raw order so assigning particular value in respective column is done in this work. Image Data: - For this dataset it required to resize the image into proper row and column. As it might possible that images in the dataset is off different size. So this is done in Pre-processing part of the image. Feature Extraction: - Here image classification is done on the basis of colour feature of the image so before encryption first it needs to be converting into grey scale. In this work if input image is in RGB format then convert grey value by $x = 0.299(R) + 0.587(G) + 0.114(B)$.

3.4 Paillier Encryption: Once the dataset get pre-process and features are extracting from it. Here client generate an encryption keys then encrypt input feature values. Finally encrypted data is send to receiver side.

3.5 Normalization & Scaling:

This step execute at server side where information is obtained in encrypted form. While server do not de-crypt the information for classification. So here this normalization step is necessitate as numbers need to convert into similar platform if it is in dissimilar level. $X = (Xi - X') / (\sigma^* \sigma)$ where X, X' denote the individual value. σ denotes mean and standard deviation. While in scaling the normalized value is multiply by a constant.

$$NS \Leftarrow \Upsilon^*X = \Upsilon^*[(Xi - X') / (\sigma^* \sigma)]$$

3.6 Kernel Function

The encrypted test sample is used to compute the polynomial kernel. $K = [(\Upsilon^*Xi)'x\ (NS) + (\Upsilon^*\Upsilon)]$. Its power is raise by p for the polynomial equation. In [8] this was done at client side while in this work same work is done at server side. This reduces the execution time of the algorithm. Now for each value obtains from the above Kernel function summation is done which help in identifying the class of the value.

3.7 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.

3.8 Proposed Algorithm at Client-End

Input: Original Data OD
Output: Encrypted data ED

```
OD[m,n]←Pre-Process(OD)
Loop 1:m
    Loop 1:n
        If OD[m, n]) != BOW
            BOW←Assign_number(BOW, OD[m, n]))
        End if
        ND[m, n] ←Intersect(BOW, OD[m, n]))
    End Loop
End Loop
Loop 1:m
    Loop 1:n
        ED[m, n]←Pailler_Algo(ND[m, n])
    End Loop
End Loop
```

Proposed Algorithm at Server_End
Input: Original Data ED
Output: Classification

```
Loop 1:m
    Loop 1:n
        NED[m,n]←Normalization(ED[m, n])
    End Loop
End Loop
K←Kernal_Calculation(NED)
Loop 1:m
    If K > Threshold
        Pos_class[m] = 1
    Else
        Pos_class[m] = 0
    End
End Loop
```

IV. EXPERIMENT AND RESULT

This section presents the experimental evaluation of the proposed Embedding and Extraction technique for privacy of image. 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.

4.1 Dataset

Experiment performed on the standard images from JAFFE database while numeric database consist known as Wisconsin Breast Cancer (WBC) values in form of Text database consist of artificial notepad files.

4.2 Evaluation Parameter

Accuracy: - Here items class after classification is evaluating. This is the percentage of correct classification done by the proposed approach.

$$Accuracy = \left(\frac{True\_class - False\_class}{Total\_class} \right) * 100$$

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.

4.3 Results

Table 1 Comparison of proposed and previous work for two class problems.

WBC Two Class Dataset Results		
Parameters	Proposed Work	Previous Work[6]
Accuracy	100	100
Execution Time	1.322	1.432

Comparison of proposed and previous work for two class problem own where both work has achieve classification accuracy of 100%. It has been obtained that proposed work has less execution time as compare to previous work. This is because as client server involvement is less.

Table 2 One to All text data classification by proposed work

Proposed Work for Text Data	
Features	Accuracy
1	66.6667
2	33.33
3	50
4	33.33

Comparison of proposed and previous work for one to all class problems is shown where proposed work has achieved high classification accuracy.

Table 3 One to All classification by proposed work for Multi class problems

Proposed and Previous Work One to All Accuracy		
Class Vs. all	Proposed	Previous
1	93.6508	66.6667
2	95.2381	69.8413
3	93.6508	66.6667
4	85.7143	53.9683

Table 4 One to All classification by proposed work for Multi class problems

Execution Time (Second) Image Data One to all		
Class Vs. all	Proposed Work	Previous Work
1	28.9023	43.8813
2	27.4309	34.531
3	24.4096	46.8713
4	28.2489	33.708

Comparison of proposed and previous work for one to all class problems is shown where proposed work has achieved better classification accuracy. It has been obtained that proposed work has less execution time as compare to previous work. This is because as client server involvement is less.

V. CONCLUSION

In this work, a set of algorithms was proposed to increase the privacy from data mining problems. As proposed work can efficiently classify all kind of data such as text, image and numeric. Here privacy was maintained by sending in encrypted form to the classifying server. The experiments showed that the proposed algorithms perform well on large databases. It is shown in the results that accuracy of the perturbed dataset is preserved for low support values as well. Here Proposed work has resolve the multi-party data distribution problem as well.

REFERENCE

[1]. Dan Boneh, Eu-Jin Goh, and Kobbi Nissim. Evaluating 2-DNF Formulas on Cipher texts. In Proceedings Of TCC 2005, Lecture Notes In Computer Science. Springer Verlag, 2005.

[2]. M. J. Abdi And D. Giveki, "Automatic Detection Of Erythema to-Squamous Diseases Using PSO–SVM Based On Association Rules", Engineering Applications Of Artificial Intelligence, Vol. 26, (2013), Pp. 603-608.

[3]. J. Nahar, T. Imam, K. S. Tickle And Y. P. Chen, "Association Rule Mining To Detect Factors Which Contribute To Heart Disease In Males And Females", Expert Systems With Applications, Vol. 40, Pp. 1086-1093, (2013).

[4]. M. Shouman, T. Turner And R. Stocker, "Applying K-Nearest Neighbor In Diagnosing Heart Disease Patients", International Conference On Knowledge Discovery (ICKD-2012), (2012).

[5]. Schulam *Et Al.,* "Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery", Associations for the Advancements of Artificial Intelligence, 2015.

[6]. 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.

[7]. H. Yu, X. Jiang, And J. Vaidya, "Privacy-Preserving SVM Using Nonlinear Kernels On Horizontally Partitioned Data," Proc. ACM Symposium Applied Computing (SAC), 2006.

[8]. 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. Transactions On Dependable And Secure Computing, Vol. 11, No. 5, September/October 2014.

[9]. 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.