International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 6, Nov-Dec.-2017, ISSN (Online): 2395-566X

A Robust Classification Algorithm for Multiple Type of Dataset

M.Tech. Scholar Afshan Idrees
Dept. of Computer Science and Engineering
Millennium Institute of Technology,
Bhopal, India
afshanidrees@gmail.com

Prof. Avinash Sharma
Dept. of Computer Science and Engineering
Millennium Institute of Technology,
Bhopal, India
avinashavi07@rediffmail.com

***Abstract* – With the increase in different internet services number of users are also increasing. Although while taking service user may be on risk for sharing data. So this work focus 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 classify all type of user data in form of text, image, numeric.**

***Keywords* – 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 clould 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 is considered to be very sensitive and must be confidential. Therefore if the little scale company thinks of using the services like classification. Classifying all account/finance related information on server makes it prone to leakage of sensitive information tell 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 futher 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, 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 diseasepatients. 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 erythemato-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.

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 fig. 1 block diagram.

Dataset: This is collection of unclassified data at the client side which is raw. So it need 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.

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 an 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.

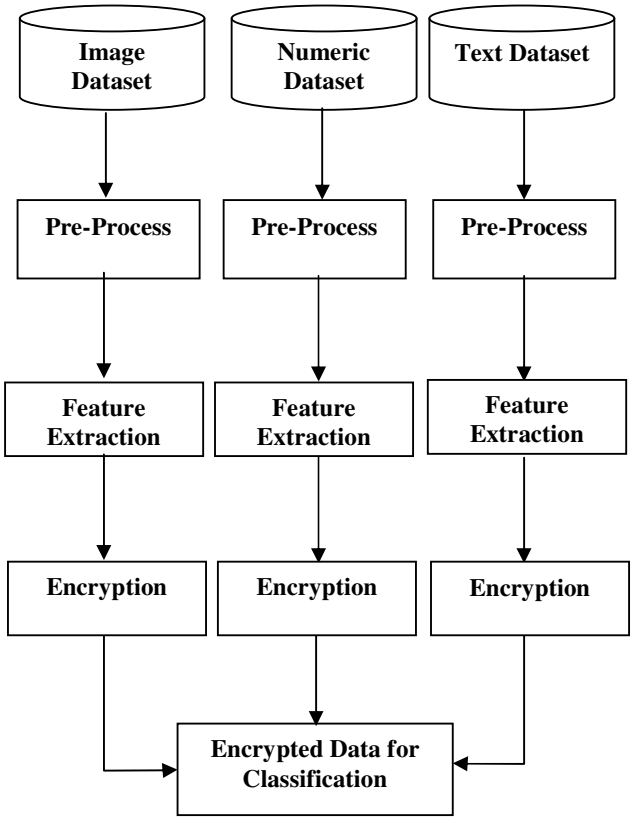

Fig. 1 Proposed Data classification model at client side.

Feature Extraction: In this step each Bag of words contain 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.

Text Encryption: So a representative of those words are required. As each keyword is a set of ASCI value for example keyword "ABCD" ASCI set is [11 12 13 14]. Now each ASCI number is replace by its binary number as 11={001011}, 12={001100}, 13={001101}, 14={001110}. So in this work ABCD binary is {00101100110000110100 1110}. Now convert this binary to decimal number for encryption.

Numeric Data: In Numeric data preprocessing 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 color feature of the image so before encryption first it need to be convert into grey scale. In this work if input image is in RGB format then convert gray value.

Advanced Encryption System: Now common step for all kind of data is that each data need to be convert into 16 element set of input. Here each input need to be in integer data type. In case of numeric this is ok, but in case of image gray scale will convert pixel values in integer form. While for text unique number is assign for all extracted words.

In this encryption algorithm four stages are perform in each round. While final round consist of three stages only. These steps are common in both encryption as well as decryption algorithm where decryption algorithm is inverse of the encryption one. So round consist of following four stages.

1. Substitute bytes
2. Shift rows
3. Mix Columns
4. Add Round Key

In final round simply all stages remain in same sequence except Mix Columns stage.

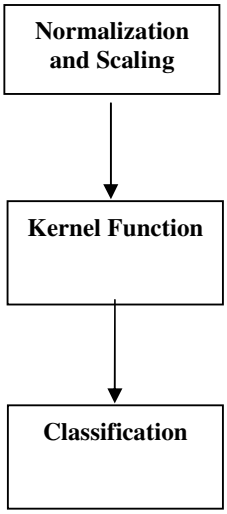

Fig. 2 Proposed Classification model at server side.

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 plateform if it is in dissimilar level.

$X = (Xi - X') / (\sigma^* \sigma)$

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

While in scalling the normalizaed value is multiply by a constant.

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

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 reduce the execution time of the algorithm. Now for each value obtain from the above Kernel function summation is done which help in identifying the class of the value.

Classification: As the decision function generate a value which is term as decision value has sign which will help in classigying the data, here the base on the positive or negative value of the decision value. Document is classify into two class.

IV. RESULTS AND ANALYSIS

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 an 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional.

4.1 Dataset: Experiment done on the standard images from JAFFE database while numeric database consist known as Wisconsin Breast Cancer (WBC) values. Text database consist of artificial notepad files.

4.2 Evaluation Parameter:

Precision = (True_positive / (False_positive+ True_positive))

Recall= (True_positive /(False_negative+ True_positive))

F-Measure = (2xPrecisionxRecall/ (Recall + Precision))

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. Precision Value for One to Many Comparison.

Image Dataset	Proposed Work	Previous Work
Set1	0.195652	0.113636
Set2	0.1875	0.130435
Set3	0.183673	0.113636

From above table 1 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption increase the precision values as well.

Table 2. Recall Value for One to Many Comparison

Image Dataset	Proposed Work	Previous Work
Set1	0.346154	0.555556
Set2	0.375	0.666667
Set3	0.391304	0.555556

From above table 2 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption increase the recall values as well.

Table 3. F-Measure Value for One to Many Comparison

Image Dataset	Proposed Work	Previous Work
Set1	0.25	0.188679
Set2	0.25	0.218182
Set3	0.25	0.188679

From above table 3 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption increase the f-measure values as well.

Table 4. Execution time for One to Many Comparison

Image Dataset	Proposed Work	Previous Work
Set1	35.9373	54.1917
Set2	35.4269	43.1663
Set3	36.1811	43.4182

From above table 4 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption reduces the execution time values as well.

Table 5. Precision Value for One to One Comparison.

Text Dataset	Proposed Work	Previous Work
Set1	0.5625	0.357143
Set2	0.642857	0.384615
Set3	0.5625	0.5

From above table 5 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption increase the precision values as well.

Table 6. Recall Value for One to One Comparison.

Image Dataset	Proposed Work	Previous Work
Set1	0.818182	0.555556
Set2	0.692308	0.555556
Set3	0.818182	0.555556

From above table 6 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption increase the recall values as well.

Table 7. F-Measure Value for One to One Comparison.

Image Dataset	Proposed Work	Previous Work
Set1	0.666667	0.434783
Set2	0.666667	0.454545
Set3	0.666667	0.526316

From above table 7 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption increase the f-measure values as well.

Table 8. Execution time for One to One Comparison.

Image Dataset	Proposed Work	Previous Work
Set1	6.52696	10.179
Set2	5.964	10.0018
Set3	6.2471	9.2356

From above table 8 it is obtained that proposed work one to many classification for image different sets is better as

compare to previous data. Here use of color feature and AES encryption reduces the execution time values as well.

Table 9. Text Data Classification Values.

Text Dataset	Precision	Recall	F-Measure	Time (Sec.)
Set1	0.428571	1	0.6	4.17821
Set2	0.375	1	0.545455	3.35648
Set3	0.428571	1	0.6	3.57644

From above table 4 it is obtained that proposed work one to many classification for image different sets is better as compare to previous data. Here use of color feature and AES encryption reduces the execution time values as well.

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.

REFERENCES

[1]. Dan Boneh, Eu-Jin Goh, And Kobbi Nissim. "Evaluating 2-DNF Formulas On Ciphertexts". *In Proceedings Of TCC 2005, Lecture Notes In Computer Science*. Springerverlag, 2005.

[2]. M. J. Abdi And D. Giveki, "Automatic Detection Of Erythemato-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 Neighbour 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 Symp. 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). 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