A Robust Privacy-Preserving Method for Outsourcing Data Classification of Numeric, Image and Text Data in Cloud

Bhavna VIshwakarma, Huma Gupta, Dr. Manish Manoria

Abstract— The persistent growth in the digital world raised privacy concern of the individual which plays a pivotal role. In cloud computing environment, the data on inter-connected network are handled by server for providing different services. In proposed work out of different cloud services emphasis is done on privacy of user data while classification. In this paper, Data Classification is performed on distinct kind of data (including numeric, image and text) in encrypted manner. For encryption, Symmetric-Key Cryptography i.e. AES (Advanced Encryption Standard) algorithm is used. On server side whole Data Classification process is executed on encrypted data only. All experiments are done on real dataset from different type of datasets. Results are compared on the basis of various evaluation parameters with previous existing methods. This paper explicates that the proposed methodology with standard algorithm efficiently classified dissimilar kind of data for distinct classes (i.e. two-class, multi-class).

Index Terms— AES, Cloud Computing, Data Classification, Data Mining, Data Outsourcing, Privacy, Symmetric-Key Cryptography.

1 INTRODUCTION

Today's digital infrastructure supports novel ways of storing, dealing out, and disseminating data. In fact, we can store our data in remote servers, access consistent and capable services provided by third parties, and use computing authority accessible at multiple locations athwart the network. Furthermore, the increasing adoption of moveable devices jointly with the dispersal of wireless relations in home and work environments has led to a more dispersed computing scenario. These advantages come at a price of superior seclusion risks and vulnerabilities as a huge amount of information is being distributed and stored, often not under the straight control of its owner. Ensuring proper seclusion and security of the information stored, converse, processed, and dispersed in the cloud as well as of the users accessing such an information is one of the big confront of our modern society.

As an issue of fact, the advancements in the Information Technology and the dispersal of novel paradigms such as Data Outsourcing and Cloud Computing, while permitting users and group to easily entrée high value submission and services, introduce novel privacy risks of indecent information revelation and dissemination. Classification is one of the usually used errands in data mining applications. For the past decade, due to the increase of diverse seclusion issues, many theoretical and practical solutions to the classification difficulty have been planned under different protection models. However, with the recent reputation of cloud computing, users now have the chance to outsource their data, in encrypted form, as well as the data mining errands to the cloud. Since the data on the cloud is in encrypted form, existing privacy preserving classification methods are not valid.

In cryptography, encryption is the method of encoding messages or information in such a way that only official revelry can read it. Encryption does not of itself avert interception, but denies the message contented to the interceptor. In an encryption scheme, the proposed statement information or message, referred to as plaintext, is encrypted using an encryption algorithm; engender cipher text that can only be read if decrypted. For technical reasons, an encryption method regularly uses a pseudorandom encryption key produce by an algorithm. It is in opinion promising to decrypt the message without possessing the key, but, for an elegant encryption method, large computational possessions and skill are necessary. An authorized beneficiary can easily decrypt the message with the key provided by the inventor to beneficiary, but does not permitted interceptors.

Use of personal data for any activity without any information to the concern is term as Public concern. In order to understand this thing consider a user never want to share their personal information with other without his permission, but it is done by the information holding organization, then this is called as public concern. Mostly it is divided into few categories such as banks; health care departments are most trustful for customer information privacy. While in case of credit card companies, some of social site they are least trust companies. Privacy preserving techniques can be classified based on the protection methods used by them.

In image classification and image retrieval, the color is the most important feature [1, 2]. The color histogram represents the most common method to extract color feature. It is regarded as the distribution of the color in the image. The efficacy of the color feature resides in the fact that is independent and insensitive to size, rotation and the zoom of the image.

Bhavna Vishwakarma *is currently pursuing Masters of Enineering degree in Computer Science & Engineering from TRUBA Institute of Engg. & IT, Bhopal (India). E-mail: bhavna.705@gmail.com.*

Huma Gupta *is currently working as Assistant Professor in Computer Science & Engineering Department and* ***Dr. Manish Manoria*** *is working as Director of Institute in TRUBA Institute of Engg. & IT, Bhopal (India) E-mail: huma.g@trubainstitute.ac.in, manishmanoria@gmail.com.*

2 Related Work

K-Nearest Neighbor (K-NN) K-Nearest Neighbor (K-NN) classifier is one of the simplest classifier that discovers the un-identified data point using the previously known data points (nearest neighbor) and classified data points according to the voting system [1]. 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 research work [2], they used predictive and priori approach for generating the rules for heart disease patients. In this 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 [3] authors 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 work [4], they 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 work [5] authors proposed a Probabilistic Sub typing 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.

3 Proposed Methodology

This proposed work is designed to classify different type of data such as Numeric, Image, and Text data. In this portion of the paper whole work is explained and all steps of proposed paradigm are shown by block diagram in fig. 1.

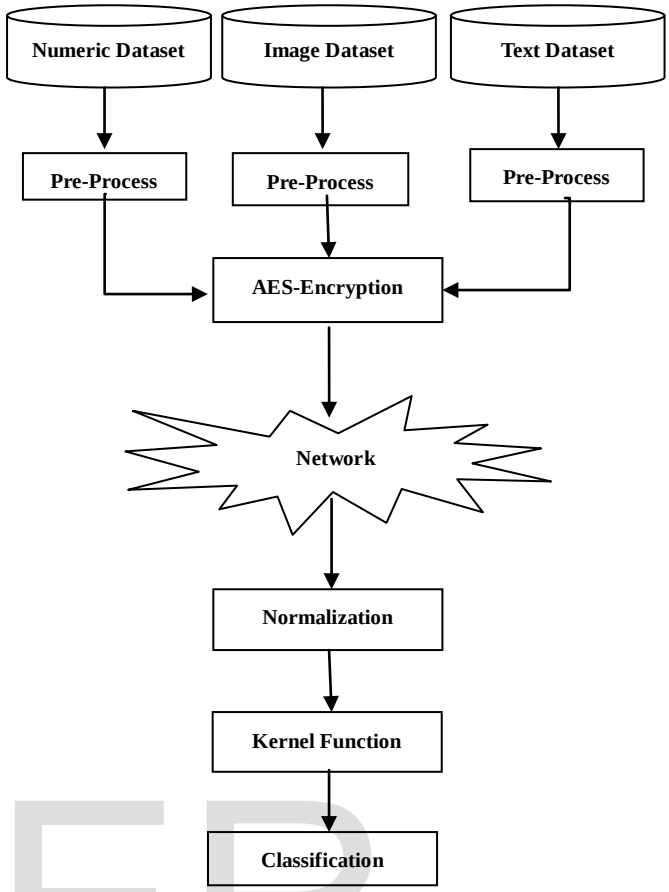

Fig. 1 Block diagram of proposed work.

2.1 Pre-Processing

In this phase different pre-processing for each numeric, image and text data is explained below.

2.1.1 Text Pre-Processing

Text pre-processing is consisting of words which are responsible for lowering the performance of learning models [7, 8, 9]. 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. 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.

This signifies that in maximum time words in corpus arises very few times in any training corpus. Those words which are arise very few times statically unimportant having low information gain. However the occurrence of any word in training in future document is very less.

The vector which contains the pre-processed data is use for collecting feature of that document. This is done by comparing the vector with vector KEY (collection of keywords) of the ontology of different area. So the refined vector will act as the feature vector for that research project/proposal.

So the lists of words which are crossing the threshold are considered as the keywords or feature of that document.

[feature] = mini_threshold ([processed_text])

2.1.2 Image Pre-Processing

In this step image is resize in fix dimension. As different image have different dimension. So conversion of each is done in this step. One more work is to convert all images in gray format. As different image has RGB, HSV, etc. format so working on single format is required.

2.1.3 Numeric Pre-Processing

In this pre-processing method proper row-column manner is used to separate different values of raw data available on data set. In this step whole data is converted into integer values and other non-numeric data is removed from the dataset.

2.2 AES Encryption

Now common step for all kind of data is that each data need to be converted into 16 element set of input. Here each input need to be in integer data type. This works fine if it's numeric, but in case of an image data gray scale will convert pixel values in integer form. While for text unique number is assign for all extracted words.

The algorithm begins with an Add Round Key stage followed rounds of four stages and a tenth round of three stages. This applies for both encryption and decryption with the exception that each stage of a round the decryption algorithm is the inverse of its counterpart in the encryption algorithm. The four stages are as follows:

- Substitute bytes
- Shift rows
- Mix Columns
- Add Round Key

The tenth round simply leaves out the Mix Columns stage.

2.3 Normalization

This step executes at server side where values obtain after encryption is normalize into fix range as some of values get much higher than other. This is very important to normalize each value as difference between the small and very large value need to be reduce. This difference arises because assigning number to text may increase gap between the text having smaller id than text having higher one.

$$normalize(t_i) = \frac{\gamma(t_i - x)}{\sigma^2}$$

Here Y is constant; ti is input value coming from client side. While X is mean and σ denote the standard deviation.

2.4 Kernel Function

In this step normalize value obtain from the above step.

$$K_t = \left| \left(C_1 \times C_2^{t_i} \right), C_3 \right|$$

Here C1, C2, C3 are constants to estimate the kernel value of ti.

In similar fashion all the values of the input vector is estimated. So sum of all these values for any text file or image or input vector of numeric data is consider as the classifying parameters.

2.5 Classification

In this step as per the summation obtained from the kernel function it lead to the decision value for estimating the class of the input vector. Here it is assumed that if summation get negative value then vector belong to one class and if summation gets positive value then vector belong to other class.

4 Experiment And Result

In this section whole experiment is explained. Here detail description about dataset with evaluation parameters is given. Then results obtained from the proposed work are compared with the existing method in [6].

4.1 Dataset

For the experiments, we considered three different datasets including two popular data sets as from the UCI machine learning repository called the Wisconsin Breast Cancer (WBC) and the JAFFE (Japanese Female Facial Expression) database. The WBC dataset is used for numeric data testing and it's about medical records of cancerous or non-cancerous patients as classified in two classes Benign & Malignant. The JAFFE is for image data testing and each subject has seven expression (or seven class) i.e., angry, disgust, fear, happy, sad, surprise & neutral face each with a pixel resolution of 256 × 256 which were resized to 51 × 51 for computational efficiency. Then we consider an artificial data set for testing text data and this artificial data set contains twelve samples and two classes. The details of the data sets are given in Table 1.

Table 1
Explaination of Different Dataset

Type	Name	Classes	Samples
Numeric	WBC	2	681
Text	Artificial	2	12
Image	JAFFE	7	213

4.2 Evaluation Parameters

- *Execution Time*

It is the time period for the algorithm to classify user data on the server. So execution time should be as less as possible. So this is a very important parameter to evaluate this work. Execution time is evaluated in terms of seconds.

- *Accuracy*

As data is classify as per the pattern in the dataset so a perfect pattern have can classify input data more perfectly then other. Accuracy is the percentage of true classification (correct class) done by the server.

4.3 Results

As shown in Table 2, it is obtained that proposed work and previous work in [6] have 100% accuracy in two class problem. While execution time required for proposed is quite less as compare to previous work because this work has reduce the client and server communication.

TABLE 2
COMPARISON OF PROPOSED AND PREVIOUS WORK FOR TWO-CLASS PROBLEMS

WBC Two-Class Dataset Results		
Parameters	**Previous Work**	**Proposed Work**
Accuracy	100	100
Execution Time	1.472	1.213

TABLE 3
ONE TO ONE CLASSIFICATION BY PROPOSED WORK FOR MULTI-CLASS PROBLEMS

Proposed Work One to One		
1 Vs Class	**Execution Time**	**Accuracy**
2	13.865	83.33
3	13.548	94.4444
4	11.874	88.8889
5	12.541	77.7778
6	10.044	88.8889
7	11.2012	88.8889

From Table 3 and 4 it is obtained that proposed work and previous work in [6] have high accuracy in multi class problem (using JAFFE dataset). While execution time required for proposed is quite less as compare to previous work because this work has reduce the client and server communication**.**

TABLE 4
ONE TO ONE CLASSIFICATION BY PREVIOUS WORK FOR MULTI CLASS PROBLEMS

Previous Work One to One		
1 Vs Class	**Execution Time**	**Accuracy**
2	23.4129	55.5556
3	23.6641	50
4	17.8834	77.7778
5	14.765	72.2222
6	15.4104	55.5556
7	15.2180	66.6667

From shown Table 5 and 6, it is obtained that proposed work and previous works in [6] have high accuracy in one to all multi-class problems (using JAFFE dataset). While execution time required for proposed is quite less as compare to previous work because this work has reduce the client and server communication.

In Table 7 it is shown that proposed work have high accuracy in one to all multi class problem for text document as well. While execution time required for proposed is quite depend on the content present in the files.

TABLE 5
ONE TO ALL CLASSIFICATION BY PROPOSED WORK FOR MULTI CLASS PROBLEMS

Proposed Work One to All		
Class Vs All	**Execution Time**	**Accuracy**
1	71.5985	77.7778
2	66.2346	76.1905
3	79.5397	73.0159
4	79.393	74.6032
5	76.491	77.7778
6	62.0621	74.6032
7	79.0043	74.6032

TABLE 6
ONE TO ALL CLASSIFICATION BY PREVIOUS WORK FOR MULTI CLASS PROBLEMS

Previous Work One to All		
Class Vs All	**Execution Time**	**Accuracy**
1	85.5010	69.8413
2	85.6766	73.0159
3	87.2287	69.8413
4	91.9482	53.9683
5	80.9063	53.9683
6	83.6942	63.4921
7	80.2102	57.1429

TABLE 7
ONE TO ALL CLASSIFICATION BY PROPOSED WORK FOR MULTI CLASS PROBLEMS (TEXT Dataset)

Proposed Work One to All		
Class Vs All	**Execution Time**	**Accuracy**
1	4.2212	58.3333
2	5.9349	66.667
3	6.6388	58.3333
4	3.8904	66.667

5 Conclusion

The privacy regarding to user data confidential information is very important. Such type of privacy may be lost during sharing of data in distributed healthcare environment. Necessary steps must be taken in order to provide proper security so that their confidential information must not be accessed by any unauthorized organizations. In this work classification of all type of data is done under encrypted form where each data from the class is encrypted first at client end then server classifies the data. It is obtained that proposed work has achieve 100 percent accuracy in two class problem, while in case of multiclass problem accuracy is highly acceptable. Experiment is done on read dataset. Results shows that proposed work is better as compare to previous existing methods. In future one can increase accuracy of multiclass by adopting some genetic algorithms.

References

[1] C. Mcgregor, C. Christina And J. Andrew, "A Process Mining Driven Framework For Clinical Guideline Improvement In Critical Care", Learning From Medical Data Streams 13th Conference On Artificial Intelligence In Medicine (LEMEDS). Http://Ceur-Ws. Org, Vol. 765, (2012).

[2] 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).

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

[4] 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.

[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] Yogachandran Rahulamathavan, Raphael C.-W. Phan, Suresh Veluru, Kanapathippillai Cumanan And Muttukrishnan Rajarajan."Privacy-Preserving Multi-Class Support Vector Machine For Outsourcing The Data Classification In Cloud ".IEEE Transaction On Dependable And Secure Computing, VOL. 11, NO. 5, September 2014.

[7] Souneil Park, Jungil Kim, Kyung Soon Lee, And Junehwa Song, Member, IEEE. Disputant Relation-Based Classification For Contrasting Opposing Views Of Contentious News Issues. IEEE transactions on knowledge and data engineering, vol. 25, no. 12, december 2013.

[8] Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu. "An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection". IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012.

[9] Ning Zhong, Yuefeng Li, And Sheng-Tang Wu "Effective Pattern Discovery For Text Mining". Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.

[10] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, And Xindong Wu "Multimodal Graph-Based Reranking For Web Image Search. Ieee Transaction On Image Processing Vol. 21, No. 11, November 2012.

[11] Wenjun Lu1, Avinash L. Varna2, And Min Wu. Confidentiality-Preserving Image Search: A Comparative Study Between Homomorphic Encryption And Distance-Preserving Randomization. Received December 15, 2013, Accepted January 15, 2014, Date Of Publication February 20, 2014, Date Of Current Version March 4, 2014.