Classification of Encryption Algorithms Based on Ciphertext Using Pattern Recognition Techniques

T. Kavitha[1], O. Rajitha[1], K. Thejaswi[1],
and Naresh Babu Muppalaneni[2](✉)

[1] Sree Vidyanikethan Engineering College (Autonomous), Tirupathi, India
[2] Department of CSE, National Institute of Technology Silchar,
Silchar, Assam, India
nareshmuppalaneni@gmail.com

Abstract. In digital era security in data communication is a challenging task. For secure data communication between the parties, encryption algorithms are being used. Recently many attacks have been reported for breaking ciphertext using various cryptanalytic techniques. Cryptanalysts are trying to break the cipher without knowing the key. Present day scenario an attacker may not know an encryption algorithm being used by communication entities. Identifying the algorithm itself is a challenging task. Once the encryption algorithm has identified, the cryptanalysts can analyzes the weakness of the encryption algorithm and be able to retrieve the plain text without knowledge of key. Here we present various pattern recognition techniques for identifying encryption algorithm which are used between two parties for communication of data using ciphertext. Thus, we consider the block cipher algorithms DES, IDEA, AES and RC operating in Electronic Code Book (ECB) mode. The classification techniques used are Support Vector Machine (SVM), Bagging (Ba), AdaBoostM1, Neural Network, Naïve Bayesian (NB), Instance Based Learning (IBL), Rotation Forest (RoFo) and Decision Trees to identify the right algorithm for the given ciphertext.

Keywords: Encryption · Ciphertext · Cryptanalysts · Pattern recognition · Block ciphers

1 Introduction

Due to a fast increase in the amount of public data transferred through the Internet has more attention for information security, of which the kernel is cryptography. The purpose of Cryptography is for storing sensitive information and for secure communication over public networks. Cryptographic algorithm protects data being attacked by attacker. Cryptanalysis is the process of analyzing and breaking of Ciphers. In Cryptanalysis if ciphertext is available two important tasks can be done, Encryption method Identification and identifying which key is used. For identifying encryption method for the given ciphertext various statistical methods and machine learning techniques can be used. The encrypted file contains Statistical methods which use the

A. P. Pandian et al. (Eds.): ICCBI 2018, LNDECT 31, pp. 540–545, 2020.
https://doi.org/10.1007/978-3-030-24643-3_64

alphabet occurrence frequency, in different methods based on machine learning techniques. Pattern classification task depends on the encryption method identification. The classification algorithms attempt to capture the inferred behaviour of every encryption method from all ciphertexts. Support vector machines are used to identify the Block ciphers which is Proposed in [1]. Methods based on machine learning are also used in crypt analysis for Feistel type block ciphers [2]. Genetic algorithm related methods are used to solve cryptographic systems; it involves finding key used for distinguishing the message [3–5]. Usage of machine learning classifiers is a growing interest of data mining [6]. Performance of classifiers is used to find user profiling.

Here we studied a number of pattern recognition techniques to identify the encryption method, and their accuracy is examined. Here ciphertext is considered as a document characterization problem, for this reason ciphertext is considered as a document and represented by document vector. Usually the ciphertext does not contain any meaning full words and symbols. Subsequence of the bit sequence in a ciphertext is represented by symbols, and sequence of these symbols is words. By using this approach, as a block cipher algorithm problem was identified by encryption method from a ciphertext. So 200 text files were considered and are encrypted with ECB (Electronic Code Book) mode. Every file is encrypted with 200 various keys.

In this paper Support Vector Machine (SVM), Naïve Bayesian (NB), Instance Based Learner (IBL), Neural Networks (MLP), Bagging (Ba), Rotation Forest (RoFo) and AdaBoost are considered and compared based Encrypted data's classification accuracy.

Generally Cryptanalysts does not have an idea for which cryptographic algorithm used when capturing ciphertext. So the cryptanalyst has to identify which cryptographic algorithm used and then decode it through brute force attack, mathematical methods, rainbow table attack and dictionary attack etc. Figure 1 shows how an identification system of cryptographic algorithm works with knowledge of only ciphertext. If ciphertext C is input into the identification system we must know the algorithm.

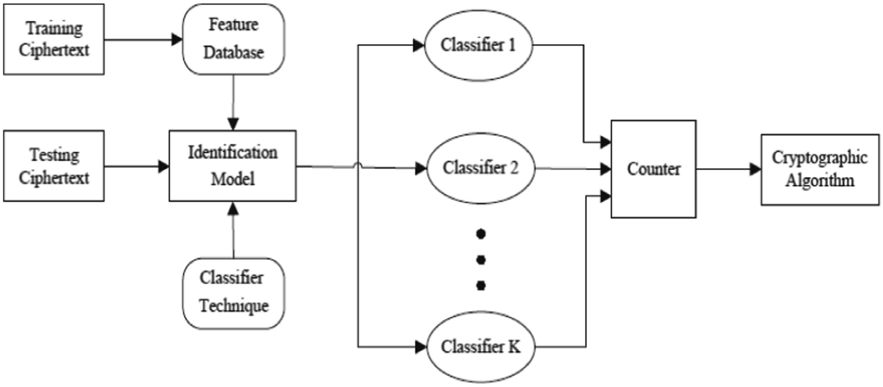

Fig. 1. Model of our identification system Cryptography Algorithm Used

1.1 Various Attacks

Ciphertext Only Attack: here only the encrypted message is available for attacker, but the language is known frequency analysis can be attempted. In this case attacker does not know anything about the contents and must work from ciphertext only.

Known Plain Text Attack: in this both plain text and ciphertext will be available in finding the key. The attacker can guess some parts of the ciphertext. Here the task is to decrypt the rest of the ciphertext blocks using this information. This can be done by determining the key used to encrypt the data.

Chosen Plain Text Attack: This occurs when the attacker gains access to target encryption device, if say it is unattended. The attacker runs various pieces of through the device for encryption. This is compared to plain text to attempt to derive the key.

Chosen Ciphertext Attack: In this the cryptanalyst can choose different ciphertexts to be decrypted and has access to it. This is generally applicable to public key cryptosystems. This involves the attacker selects certain ciphertexts to be decrypted, and then using the results of these decryptions to select subsequent ciphertexts.

2 Methodology

Naive Bayesian: The naive Bayesian classifier is the most straight forward and computationally efficient classifier. It is being used as a best classifier, and it needs small amount training data to estimate parameters. If structure is known then easy to construct and one more advantage is classification process is efficient [7]. This leads to assumption that all features are independent with each other.

$$\mathrm{Similarity}\ (\mathrm{a,b}) = -\sqrt{\sum_{i=1}^{n} f(a_i, b_i)}$$

Support Vector Machines: support vector machines are learning algorithms that analyze the data for classification and regression, it performs linear classification in addition to this it efficiently performs non linear classification. SVM algorithms have been widely applied text and hypertext classification and classification of images and hand written characters are also recognized using SVM's.

Neural Networks: neural network consists of input layer, output layer and hidden layers. Here neuron is taken as a adaptive element with weights, its value modifies based on inputs and outputs the mathematical function is network is tested for random texts with random keys of various lengths [9].

Instance Based Learner: By using specific instances IBL generates classification prediction. Unlike nearest neighbour algorithm, this method normalizes its processes instances, attributes ranges, incrementally and it has a policy for tolerating missing values [10]. The instance based learners applies Euclidean distance function to supply grade matches between given test instance and training instances [11].

Bagging: Bagging is the application of the Bootstrap procedure for machine learning algorithm, typically decision trees. it is a ensemble method that generates individuals by training each classifiers with random redistribution by taking original data with size n, these is known as bootstrap replication of original set. Bootstrap replicate contains 60% of the original set [8, 12].

AdaBoost Algorithms: it is a ensemble algorithm, the idea is create strong classifier from weak classifiers. In this approach each classifier is enough to moderately accurate, this is better than random guessing. In this approach each weak classifier is added in a different level so that it will classify the data. Set of weights is assigned to objects in the data set, so that difficult objects acquire more weight.

Rotation Forest: this algorithm is used for feature extraction and is based on Principal Component Analysis (PCA). Principal component analysis is applied on base classifier by splitting the feature set into k subsets. Here variability information is preserved hence k axis rotation is takes place on base classifier [8].

3 Encryption Method Identification

In this paper we are building a model to predict the block cipher encryption method from the given ciphertext. The algorithms considered are IDEA, DES, AES, and RC2. The block sizes of these algorithms are DES (56-bits), IDEA (64-bits), AES (128, 192, 256 bits) and RC2 (42, 84, 128 bits). 200 text files of fixed size (512 KB) have been used for this study. Bouncy Castel Crypto API library [6] was used to achieve encryption. All these files are encrypted with 200 dissimilar keys to make more difficult to identify. Simulation and Performance Evaluation for identification of cryptographic algorithm mainly focus on some experiments. 200 plaintext and ciphertext pairs were used for this study. For each cryptographic algorithm, we have taken input training ciphertexts as 40 and remaining are divided into 8 groups of 20 ciphertexts for testing. The process of encryption runs in ECB mode. Here for the same cryptographic algorithm the key is fixed. It is clearly observed for final identification of algorithm the size of ciphertext and key plays major role. The following table shows the same key is used for training and testing ciphertext.

Here two different Data Sets are considered; the set P contains 8 classes with different key for different encryption algorithms and 220 input files. The another dataset called Q contains 4 classes for each algorithm and 220 input files. The data are mainly divided into 10 partitions, 9 out of 10 data is used for training and one out of 10 for testing. The process is repeated 12 times; here the overall error rate is reduced. Using the same training set all the classifiers were trained and have been tested for making the classification accuracy.

The simulation results show how accurately classification is done with classifiers. That is Dataset P and Dataset Q. for Dataset P is used for classifiers. In the Table 1, it is observed that RoFo achieves a high accuracy classification (31.90). IBL has the least accuracy classification (11.50) means that only 28 input data out of 220 were classifies correctly.

Table 1. Classification of accuracy of the classifiers

Classifiers	Accuracy	
	Dataset P 8 Classes	Dataset Q 4 Classes
SVM	17.08	32.08
BA	25.83	48.7
AdaBoostM1	20.42	42.23
NN	24.25	45.87
NB	28.33	44.17
IBL	11.50	30.20
RoFo	**31.90**	**54.32**
Decision trees	27.25	43.06

The simulation results of Dataset –Q shows RoFo again performs well than all other classifiers with the accuracy of (54.32) means 130 input data of 220 were classified. Once again the IBL has least classification accuracy (30.20) means only 73 input data are classified.

4 Conclusion

The goal is to determine the best classification algorithm with high accuracy for these four Block ciphers IDEA, DES, AES and RC2 are considered. In this paper eight classifiers (NN, BA, AdaBoostM1 and SVM, NB, IBL, RoFo, and Decision trees) are used for identifying encryption method and its accuracy are evaluated. These simulations are performed using WEKA tool. The different keys are used for different text data. The simulation results shows Rotation Forest (RoFo) classifier has highest classification accuracy for identifying encryption method for ciphered data, where as IBL has least Performance. The accuracy is only 54%, hence new techniques has to develop for identification of encryption method.

References

1. Girish, C.: Classification of modern ciphers. M. Tech thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur (2002)
2. Brahmaji, M.: Classification of RSA and idea ciphers. M. Tech thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur (2003)
3. Saxena, G.: Classification of ciphers using machine learning. M. Tech thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur (2008)
4. Dileep, A.D., Sekhar, C.C.: Identification of block ciphers using support vector machines. In: Proceedings of the International Joint Conference on Neural Networks, Vancouver, BC, Canada, pp. 2696–2701 (2006)

5. Nagireddy, S.: A pattern recognition approach to block cipher identification. M. Tech thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Madras (2008)
6. Mishra, S., Bhattacharjya, A.: Pattern analysis of ciphertext: a combined approach. In: Proceedings of the International Conference on Recent Trends in Information Technology, pp. 393–398 (2013)
7. Chopra, J., Satav, S.: Impact of encryption techniques on classification algorithm for privacy preservation of data. Int. J. Innov. Res. Sci. Eng. Technol. **2**(10), 5398–5402 (2013)
8. Cufoglu, A., Lohi, M., Madani, K.: Classification accuracy performance of naïve Bayesian (NB), Bayesian networks (BN), lazy learning of Bayesian rules (LBR) and instance-based learner (IB1)-comparative study. In: 2008 International Conference on Computer Engineering and Systems, pp. 210–215 (2008)
9. Kuncheva, L.I., Rodríguez, J.J.: Classifier ensembles for fMRI data analysis: an experiment. Magn. Reson. Imaging **28**, 583–593 (2010)
10. Maeda, Y., Wakamura, M.: Simultaneous perturbation learning rule for recurrent neural networks and its FPGA implementation. IEEE Trans. Neural Netw. Publ. IEEE Neural Netw. Counc. **16**, 1664–1672 (2005)
11. Muchoney, D., Williamson, J.: A Gaussian adaptive resonance theory neural network classification algorithm applied to supervised land cover mapping using multitemporal vegetation index data. Network **39**, 1969–1977 (2001)
12. Harvey, I.: Cipher Hunting: How to Find Cryptographic Algorithms in Large Binaries. N cipher Corporation Ltd., Cambridge (2001)