Analysis of cryptosystem recognition scheme based on Euclidean distance feature extraction in three machine learning classifiers

SiJie Fan¹, YaQunZhao² and SiJie Fan*¹

¹State Key Laboratory of Mathematical Engineering and Advanced Computing, 450001, China
²State Key Laboratory of Cryptography and Science, 450001, China
*Corresponding author’s e-mail: SiJie Fan; email: 826733148@qq.com

Abstract. In reality, many cryptographic analysis techniques are based on a specific cryptographic system or a large number of encrypted ciphertext. The identification and detection of cryptographic system is of great significance for evaluating the security of the algorithm and guiding the design and improvement of the algorithm. In this paper, we transcode each character in ciphertext into a decimal number, construct these numbers into one-dimensional arrays, and obtain the Euclidean distance between these one-dimensional arrays. Then we use these distances as features and input them into three machine learning classifiers: random forest, logistic regression and support vector machine to recognize cryptosystem and compare their recognition accuracy. The subjects include 8 common block ciphers (DES, 3DES, AES-128, AES-256, IDEA, SMS4, Blowfish, Camellia-128). The experimental results show that using the feature extraction scheme not only shortens the experimental time, reduces the computational cost, but also improves the recognition accuracy of eight typical block cipher algorithms. The classification accuracy of the ECB mode in the random forest classifier is 75%, which is higher than the existing published literature experimental results. The classification accuracy rate of CBC mode is higher than 13.5%, which is higher than the accuracy of random classification.

1. Introduction
Usually we call the process of distinguishing and identifying the original information encryption algorithm as cryptosystem recognition. In modern cryptography, cryptanalysis is usually carried out on the premise of a large number of ciphertext obtained. It is of great significance to classify and detect cryptosystem for evaluating the security of cryptosystem. At present, the classification method based on machine learning is the main method of cryptosystem recognition[1]. The cryptosystem recognition scheme based on machine learning regards the cryptographic system as a classification problem, and the extracted feature values are tagged into a neural network classifier for classification. The cryptosystem recognition scheme based on machine learning technology has remarkable effect. It is simple in design, stable in performance and can deal with the complex relationship between data. Therefore, the prospect is very promising. Nowadays, many machine learning algorithms are applied to cryptosystem recognition schemes.

In 2006, Dileep [2] and others proposed a block cipher recognition scheme based on support vector machine (SVM) with the help of text categorization and counting. This paper compared the
recognition performance of SVM and K-nearest neighbor methods, and adopt multi-cryptosystem including fixed-length document vector and variable-length document vector. In this article, the author mainly discussed the distinction between five cryptosystems including AES, DES, 3DES, Blowfish, and RC5, and studied the cryptosystem identification under different conditions such as fixed key and variable key, ECB mode and CBC mode. In 2008, Nagileddy et al. [3] considered that cryptosystem identification was a preliminary cryptographic attack and can be used to evaluate the security of a cryptosystem. At the same time, the literature also identified five cryptosystems such as AES, DES, 3DES, Blowfish and RC5. The conclusions showed that the block cipher system in ECB mode was easier to identify, and the CBC mode can resist this cryptosystem identification. At the same time, AES had lower recognition accuracy than other cryptosystems. In the same period, Caurav Saxena et al. gave a distinction between classical and modern cryptosystems, and carried out recognition experiments on cryptosystems such as Blowfish, Camellia and RC4 through support vector machines and other pattern methods. According to the results of the predecessors, it can be found that different classification techniques had different ability to distinguish various algorithms. In 2010, Suhaila et al. [4] [5] used pattern recognition method to compare cryptosystem recognition classifiers constructed by eight classification algorithms (Naive Bayesian, Support Vector Machine, Neural Network, Instance Based Learning, Bagging, AdaBoostM1, Rotation Forest and C4.5 Decision Tree), block cipher algorithms in four ECB modes (56-bit DES, 64-bit IDEA, 128, 192, 256-bit AES and 42, 84, 128-bit RC2) was used as recognition objects. The experimental results showed that the RoFo classification algorithm had a relatively high classification accuracy rate, and the success rate of the 8 types of data sets reached 30.83%. It can successfully distinguish 74 out of 240 files, and can distinguish 240 in 4 types of data sets. In the four types of data sets, 128 of the 240 can be distinguished, the success rate was 53.33%, and the IBL had the worst performance, with only 12.5% success rate for 8 data sets and 30.42% for 4 data sets.

All of the above methods used binary analysis code, which often required reverse analysis of the encryption program, resulting in its poor applicability in non-cooperative application environment. In this paper, we propose a feature extraction scheme based on Euclidean distance. The ciphertext encrypted by eight block cipher algorithms (DES, 3DES, AES-128, AES-256, IDEA, SMS4, Blowfish, Camellia-128) is used as training. After that, we use three machine learning classifiers to conduct experiments and compare the experimental results. Firstly, we find that random forests have better classification effect based on the same training data. Secondly, we use the feature extraction scheme based on Euclidean distance proposed by us, which greatly reduces the dimension of feature vectors, reduces the scale of experiments and reduces the running time of experiments. Compared with the previous work, the accuracy of this experiment has been greatly improved.

2. Background knowledge

2.1. Brief Introduction to Block Ciphers

Block cipher is one of the important systems in modern cryptography[6], which is an important part of many cryptosystems. As a basic construct, the versatility of block ciphers can also constitute components such as random number generation, sequence ciphers, message authentication, and hash functions[7]. Block cipher usually refers to a kind of cipher algorithm that can only deal with a piece of data of a certain length at a time. Here, the "piece" is called a block. The number of bits in a block is called the block length. Specifically, the principle of block cipher is to divide the plaintext message sequence $m_1, m_2, \ldots, m_n$ into a group $(m_1, m_2, \ldots, m_n), (m_{n+1}, m_{n+2}, \ldots, m_{2n}), \ldots$, encrypts it according to a set of fixed encryption algorithms under the control of the key $K = k_1, k_2, \ldots, k_n$, and outputs a group of ciphertext $(c_1, c_2, \ldots, c_n), (c_{n+1}, c_{n+2}, \ldots, c_{2n}), \ldots$. The model is shown in the figure1.
2.2. Two working modes of block cipher

The ECB (Electronic Codebook) mode is the simplest encryption mode, which divides plaintext messages into fixed-length blocks. Each block is individually encrypted. All blocks use the same method for encryption and decryption of encryption and are independent of each other, so parallel computing can be performed[9].

2.3. Common Distance Measurement Methods in Machine Learning

Euclidean distance refers to the distance between two points in n-dimensional space.
3. Experiments and Result Demonstration

During the data collection phase, we use the data from the Caltech-256 dataset[8] of the California Institute of Technology to convert the data into 1001 files for text stitching. Each of the eight cryptosystems contains 1000 sub-files, each of which has a size of about 513 KB, and then we encrypt them with the 16 encryption algorithms mentioned above to obtain 1001 text files. Then we cut off these ciphertext files, the file length is still controlled at 512KB. Finally, there are a total of 16016 ciphertext files, which are 512 KB in size, including 8008 ciphertext files, of which 8 algorithms are in ECB mode and 8008 ciphertext files are in CBC mode. We use the open source tool OpenSSL to encrypt the plaintext, encrypt the plaintext with the same key during the training phase and the testing phase.

3.1. Experiment procedure

3.1.1. Feature extraction

1) Transcode the ciphertext file to display as ASCII code, convert all ASCII codes into decimal numbers, each 256 ASCII codes represents a group, a total of \[ \left\lfloor \frac{n}{256} \right\rfloor \] groups can be obtained, \( n \) is the dimension of each group, at this time ciphertext can be expressed as a matrix

\[
Cipher = [c_1, c_2, \cdots, c_{\left\lfloor \frac{n}{256} \right\rfloor}] = [[[c_{1,1}, c_{1,2}, \cdots, c_{1,256}]], [c_{2,1}, c_{2,2}, \cdots, c_{2,256}], \cdots, [c_{\left\lfloor \frac{n}{256} \right\rfloor,1}, c_{\left\lfloor \frac{n}{256} \right\rfloor,2}, \cdots, c_{\left\lfloor \frac{n}{256} \right\rfloor,256]]
\]

2) Transpose the matrix to get the ciphertext matrix after disturbing the original sequence.

\[
Cipher'=[c_1', c_2', \cdots, c_{256}'] = [[[c_{1,1}', c_{1,2}', \cdots, c_{1,256}']], [c_{2,1}', c_{2,2}', \cdots, c_{2,256}'], \cdots, [c_{\left\lfloor \frac{n}{256} \right\rfloor,1}'}, c_{\left\lfloor \frac{n}{256} \right\rfloor,2}', \cdots, c_{\left\lfloor \frac{n}{256} \right\rfloor,256}']]
\]

3) Compute the Euclidean Distance of All Adjacent Groups of the New Ciphertext Sequence Matrix

\[
d = [d_1, d_2, \cdots, d_{255}], \quad d_i = \sqrt{\sum_{j=1}^{256} (c_{j,i} - c_{j,i+1})^2}, \quad i = 1, 2, \cdots, 255
\]

Then standardize it as a feature of the ciphertext.

3.1.2. Building a classifier model

We paste the extracted feature vector with the encryption algorithm tag of the file source it belongs to. The labeled experimental data are divided into training set and test set by cross-over method. The training set and test set are input into Machine learning classifier, and the total accuracy is equal to 10 average values. The human interference is eliminated to the maximum extent, and the rationality of the model is determined. The specific process is shown in Figure 3 below.
3.2. Experimental results

3.2.1. Classification results of cryptosystem based on random forest
Random forest is composed of several CART (Classification and Regression Tree). For each tree, the training set they use is sampled from the total training set. When training the nodes of each tree, the features they use are randomly extracted from all the features according to a certain proportion without playback. In this experiment, the ratio is 1/2 sqrt (255*1001). The training process is shown below.

1) Given training set S, test set T and feature dimension F. Determine parameters: the number of CARTs used t, the depth of each tree d, the number of features used by each node f, termination conditions: the minimum number of samples s on the node, the minimum information gain m on the node, for the (1-t)th tree, i = 1−t.

2) From S, there is a training set S(i) with the same size as S, and as a sample of the root node, training is started from the root node.

3) If the termination condition is reached on the current node, the current node is set as a leaf node. In the classification problem, the predicted output of the leaf node is the most c(j) of the current node sample set, and the probability p is c(j). The proportion of the current sample set. If the current node does not reach the termination condition, the f-dimensional feature is randomly selected from the F-dimensional features without being put back. Using this f-dimensional feature, the one-dimensional feature k with the best classification effect and its threshold θ are searched. The sample with the k-th dimension of the sample on the current node less than θ is divided into the left node, and the rest is divided into the right node. Continue to train other nodes.

The experimental results are as follows:

| Number of experiments | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Recognition accuracy  | 0.763 | 0.742 | 0.726 | 0.772 | 0.766 | 0.723 | 0.753 | 0.764 | 0.738 | 0.751 |

Table 1. Experimental results based on random forest classification
3.2.2. Classification results of cryptosystem based on SVM

The support vector machine method is based on the VC dimension theory and structural risk minimization principle of statistical learning theory. According to the limited sample information, the complexity of the model and learning ability in order to get the best promotion ability (or generalization ability). The experimental process can be simplified by selecting two alphas in each iteration (where alpha is a hyperparameter that can be adjusted to get the optimal results) for optimization. Once you find a suitable pair of alpha, increase one and decrease the other. The "suitable" here means that the two alphas must meet certain conditions. The first condition is that the two alphas must be outside the interval boundary. The second condition is that the two alphas have not been intervalized. Or not on the border. The experimental results are as follows.

| Number of experiments | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Recognition accuracy  | 0.407 | 0.412 | 0.397 | 0.396 | 0.396 | 0.410 | 0.368 | 0.344 | 0.398 | 0.353 |

Figure 5. Histogram of experimental results based on random forest classification

Figure 6. Histogram of experimental results based on SVM classification
3.2.3. Classification results of cryptosystem based on logistic regression

In multi-category logistic regression, dependent variables are predicted by a series of independent variables (i.e., the characteristics and observation variables we call them). Specifically, after linear combination of independent variables and corresponding parameters, a probability model is used to calculate the probability of obtaining a certain result in the predictive dependent variables. The parameters corresponding to independent variables are calculated by training data. Sometimes these parameters become regression coefficients. The experimental steps can be summarized as follows: training N classifiers with one class as a positive example and other classes as a negative example. If only one classifier is predicted to be positive in the test, the corresponding classes are marked as the final classification results. If more than one classifier is predicted as an example, the data after logical regression is used as the criterion to select the maximum value. The experimental results are as follows:

| Number of experiments | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----------------------|---|---|---|---|---|---|---|---|---|----|
| Recognition accuracy  | 0.610 | 0.633 | 0.639 | 0.616 | 0.655 | 0.653 | 0.650 | 0.647 | 0.618 | 0.648 |

Figure 7. Histogram of experimental results based on logistic regression classification

4. Conclusions

By comparing the results of previous work, we get the following table 4.

| Sources of experimental results | This paper | [1] | [2] | [5] | [14] |
|-------------------------------|-----------|----|----|----|------|
| Number of Cryptosystem Types  | 8         | 5  | 5  | 8  | 10   |
| ECB mode classification accuracy | 75%       | 21.5% | 41% | 30.84% | 36.65% |
| CBC mode classification accuracy | 13.5%     | 20% | 20% | 12.5% | 20%  |

From the above table, we can see that our cryptosystem recognition scheme has higher recognition accuracy and can support more kinds of cryptosystem recognition tasks. The recognition accuracy rate...
in ECB mode is 75%, which is higher than other existing work. The recognition accuracy rate in CBC mode is 13.5%, which is higher than that of random classification.

In future research, we will design better classification features based on different working modes of block cipher encryption algorithm to identify current slow-moving cryptosystem schemes, such as CBC mode. At the same time, we can also study the stream cipher and the public key cipher to increase the applicability, robustness and universality of the scheme.

Acknowledgments
This work is supported by National Key Research and Development Project 2016-2018 (2016YFE0100600); State Key Laboratory of Information Assurance Technology Open Fund Project (KJ-15-008); State Key Laboratory of Cryptography and Science.

References
[1] Huang Liangwei, Zhao Zhicheng, & Zhao Yaqun. (2018). Hierarchical identification scheme for cryptosystem based on random forest. Journal of Computers.
[2] Dileep, A. D., & Sekhar, C. C. (2006, July). Identification of block ciphers using support vector machines. In The 2006 IEEE International Joint Conference on Neural Network Proceedings (pp. 2696-2701). IEEE.
[3] Nagireddy, S. (2008). A pattern recognition approach to block cipher identification. Indian Institute of Technology Madras.
[4] Sharif, S. O., & Mansoor, S. P. (2011). Performance evaluation of classifiers used for identification of encryption algorithms. ACEEE Int. J. on Network Security, 2(4), 42-45.
[5] Sharif, S. O., Kuncheva, L. I., & Mansoor, S. P. (2010, December). Classifying encryption algorithms using pattern recognition techniques. In 2010 IEEE International Conference on Information Theory and Information Security (pp. 1168-1172). IEEE.
[6] De Canniere, C., Biryukov, A., & Preneel, B. (2006). An introduction to block cipher cryptanalysis. Proceedings of the IEEE, 94(2), 346-356.
[7] Ren Wei. (2011). Cryptography and Modern Cryptography Research. Information Network Security, (8), 0-0.
[8] Gu Lize, Zheng Shihui, & Yang Yixian. (2009). Modern Cryptography Course. Beijing University of Posts and Telecommunications Press.
[9] Wu Wenling, & Feng Dengguo. (2006). Research status of block cipher working mode (Doctoral dissertation).
[10] Forest, R. (2003). A Classification and Regression Tool for Compound Classification and QSAR Modeling Svetnik, Vladimir; Liaw, Andy; Tong, Christopher; Culberson, J. Christopher; Sheridan, Robert P.; Feuston, Bradley P. Journal of Chemical Information and Computer Sciences, 43(6), 1947-1958.
[11] Xuegong, Z. (2000). Introduction to statistical learning theory and support vector machines. Acta Automatica Sinica, 26(1), 32-42.
[12] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.
[13] Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset.
[14] Palit, S., Sinha, S. N., Molla, M. A., Khanra, A., & Kule, M. (2011, September). A cryptanalytic attack on the knapsack cryptosystem using binary firefly algorithm. In 2011 2nd International conference on computer and communication technology (ICCCT-2011) (pp. 428-432). IEEE.