An overview of profiling side channel analysis based on machine learning

Shijie Song1*, Kaiyan Chen1 and Jiawei Ma1
1Army Engineering University, Shi Jia Zhuang, HeiBei, 050051, China
*Corresponding author’s e-mail: ssj710640965@163.com

Abstract: It has made a real issue in cryptographic security with the widespread use of them which attracts more attention now. This paper summarizes the most distinguishing profiling type attack in the side channel of cryptographic chips. The principle steps of the traditional profiling side channel analysis are introduced firstly, then the exposed problems are directed in discussion. It fits well to solve the problems using the algorithm in machine learning. So what undoubtedly added in is that the profiling side channel analysis is based on machine learning. Also the situation and the progress are both the significance in the paper. The disadvantages and the future of its development are mentioned at the end.

1. Introduction
In 1996, Paul Kocher [1] proposed the side channel analysis (SCA) the first time, which broke the traditional cryptanalysis research ideas. Bypassing the research on the mathematical security of cryptographic algorithms, it directly analyzes the particular equipments. The cryptographic operations leak some physical information related to the keys, which can be collected to obtain the key of the related devices. As a result, the information has been stolen while the traditional password analysis seems safe.

SCA can be divided into two categories: profiling SCA and non-profiling SCA. Profiling SCA aims to copy a template to implement a SCA. Generally, it is called profiling SCA while it also stands for the traditional Gaussian model-based template attack [2] and stochastic attacks [3] based on linear regression. Non-profiling SCA always uses statistical models to analyze, including differential power analysis (DPA) [1], correlation power analysis (CPA) [4], collision analysis (CA) [5], and mutual information analysis (MIA) [6] and so on.

2. Traditional profiling SCA
The traditional template SCA is to obtain the same equipment as the target device, and to model the collected signals through feature engineering then implement matching attacks. In theory, the profiling SCA is the best SCA method under the premise of obtaining the same target device.

The idea of profiling SCA is also based on the statistical analysis of the assumptions of intermediate values in the encryption phase. The adversary records the leakage \( l_{x,k} \) measured on the cryptographic device, where the \( x \) is the known plain text (bytes), and the \( k \) is the target keys (bytes). During the attack, the adversary uses a series of template leaked signal \( l \) for profiling first. Then it uses the candidate model parameter which is the probability density function of \( \theta_{x,k} \) to calculate the probability, and the adversity obtains a new attack trajectory to solve the maximum a posteriori probability of the key \( k \). As the formula (2-1) goes:
There are three profiling SCA based on the above theory of mathematics, including feature extraction, template construction, and template matching.

(1) Feature Extraction:

Feature extraction and splitting is the preprocessing process of feature engineering. The position of the feature points extracted by the SOD [2] (Sum Of pairwise Differences) method was proposed by Chari et al. Christian [7] and others proposed an improved method SOSD, adding a square operation based on the SOD method to solve the problem of the difference sign. However, the effect of SOD and SOSD will be significantly reduced when the number of samples is limited and data of sufficient scale cannot be collected. Gierlichs [8] et al. proposed the SOST algorithm based on the above problems, and introduced T-test to eliminate the statistical impact of sample size on mean and variance.

(2) Template Construction:

In the previous mathematical theory of traditional template SCA, we can choose a Gaussian template or a stochastic template for the undetermined model parameters \( \theta_{x,k} \). It means that the parameter \( \theta_{x,k} \) corresponds to the mean vector \( \mu_{x,k} \) and the covariance matrix \( \Sigma_{x,k} \) by using the Gaussian model estimation, and the mean vector and the covariance matrix of the best template are obtained by the maximum likelihood principle to find the key. Using a stochastic model, at the same time, the data operation part \( h_i(x,k) \) and the zero-mean noise part \( R_i \) are divided into the measured trajectory \( I(x,k) \), and \( I_i(x,k) \) and \( h_i(x,k) \) The differences of the two are multi-time accumulated to represent the noise covariance, and then the maximum probability value is obtained by the maximum likelihood principle, which is an unknown key. As the formula (2-2):

\[
\alpha(x_1, \ldots, x_N; k) = \prod_{j=1}^{N} f_0(\hat{h}_i(x_j, k^*)) - h_i^*(x_j, k)) \tag{2-2}
\]

\( f_0 \) represents the covariance matrix, \( k^* \) represents the unknown correct key, and \( h_i^*(\cdot) \) represents the estimated value of \( h_i(\cdot) \). For the comparison between TA and SA (Stochastic attack), Gierlichs et al. [8] confirmed that the success rate of TA is higher when the amount of feature point data is larger, and SA is more accurate when the amount of analysis data is relatively small. This also proves the applicability of TA to high-dimensional large data volume.

(3) Template Matching

The classical random sample matching methods in multivariate statistics mainly include distance discrimination, Bayesian discrimination and Fisher discriminant [10]. The distance discrimination includes statistical distance and Mahalanobis distance. In addition, we can also match the classic CPA and DPA methods without profiling SCA in the matching phase, and the difference is that the analysis efficiency is higher than the template with the original method.

3. Profiling Side Channel Analysis based on Machine Learning

Most of the traditional profiling side channel analyses use the Gaussian model, which cause the cumbersome data dimension explosion and classification process while Machine Learning [11] can be the optimal alternative.

According to the learning styles, ML technology can be divided into three categories: unsupervised learning [12], semi-supervised learning [13] and supervised learning [14]. Unsupervised learning is mainly used when information (training data sets) is not available and does not require any prior verification or data modeling. Typical examples are clustering (eg K-means [15]) and dimensionality reduction (eg principal component analysis PCA [16]). Supervised learning refers to the techniques involved in modeling a training data set (marked data set), including classification and regression.
Once the learning starts, a supervised learning algorithm is executed, and the algorithm returns the most accurate output for the new input based on the previously learned model. The main methods currently used in supervised machine learning for template attacks include: neural network (NN) [17], random forest (RF) [18], support vector machine (SVM) [19] and so on. Semi-supervised learning relies on hypotheses for direct push and induction, which combines the characteristics of both of these learning styles: Training a large number of unlabeled samples with limited-labeled samples. It avoids the waste of data, and also avoids the disadvantages of weak supervisory generalization ability and low unsupervised learning accuracy.

Supervised learning learns new data through trained tags, which corresponds to the modeling phase of the profiling SCA that the process of feature extraction and classification. In combination with the traditional template attack, the following steps are taken to implement the attack:

1) Dimensionality reduction. The preprocessing step for raw data usually refers to dimensionality reduction. The number of trajectories we obtained in the actual experiment is very large, so the relevant leakage feature points are selected for the intermediate value of a single key correlation for preprocessing. The linear discriminant analysis LDA and principal component analysis PCA are usually used for dimensionality reduction.

2) Data set classification. In order to make efficient use of the data, the dimensionally processed data will be divided into training data sets and test data sets so that machine learning can be tested after modeling.

3) Training. Training is the learning process that supervises learning. It selects the appropriate model with the right place. Some common training models are: Support Vector Machine (SVM), Neural Network (NN), Decision Tree (DT), Random Forest (RF), and \( k \)NN [20]. Those can replace Gaussian or stochastic models in traditional template attacks.

   ① Support Vector Machine (SVM): The support vector machine is a classifier that supports vector operations. The algorithm can determine the classification boundary by using only part of the data composition vector, that is, constructing the optimal separation hyperplane to maximize the distance (the boundary) from the nearest data point of each class. It may happen that these classes are not linearly separable or have some constant. Therefore, applying kernel techniques [21] is a method of creating non-linearly separated surfaces. The largest boundary plane may still be linear in the transformed feature space. The kernel has many parameters to choose from. For example, kernel functions can be linear, polynomial, radial basis, or sigmoid. The most common function is the radial basis function (RBF). Therefore, SVM technology has a good classification effect on nonlinear classification problems. The support vector machine method is recognized as the most effective and research use in profiling SCA with its excellent generalization ability and high learning precision.

   According to the characteristics of SVM technology, Hospodar and Gierlichs [22] successfully optimized the encryption intermediate value analysis by using LS-SVM technology on block cipher software for the first time. Heuser and Zohner [23] created a general methodology for SVM attacks and compared this method with a template attack, which proved that SVM is better than TA. In 2013, Bartkewitz and Lemke-Rust [24] applied a multivariate SVM machine learning model to improve the attack success rate compared to the single-bit method. In addition, they use (linear) SVM as a preprocessing tool for feature selection, similar to the method of Brank [25]. Then, for the protected encryption algorithm, Lerman [26] and others used SVM to perform a template attack on AES with a rotating S-box mask and compared it with TA and SA. The results show that using the support vector machine only needs 26 measurements and retrieval to complete the crack, which is better than other algorithms. Another attack based on support vector machine (SVM) was proposed in [27]. The author uses a support vector machine to recover the key bit by bit by using the key in the key replacement loop.

   ② Neural Network (NN): The perceptron of the neuron-like cell mechanism is the simplest neural network model in which the input values are accumulated by weights and the classification tags are output through the activation function. Multiple Layer Neural networks are combined by a perceptron into a complex structure to process the data to be classified. The neural network (NN) is mainly used for branches of the cryptographic to implement key distribution [28], hash function [29], random number
generator [30], public key cipher [31] and exchange protocol [32]. A power analysis method based on Multiple Layer Perceptron was first proposed in [33]. In this work, the author uses the simple MLP in the neural network to attack the AES key, and can determine the first byte of the key with a probability of about 90%, which can achieve single-digit trajectory cracking. In [34], the MLP method pre-processed by using the measured power trajectory is pre-processed, the optimization is not computationally demanding, the probability of suppressing the corresponding key estimation is reduced, and the reduction may tend to be wrong. The number of keys categorized makes the estimated success rate nearly 100%.

Decision Tree (DT) and Random Forest (RF): Each internal node of DT tests an attribute (such as the average). From top to bottom, each branch corresponds to an attribute value, and each leaf is assigned to a classification. After that, the training data is organized into a tree, and the classification of the new trajectory is equivalent to the tree searching and the decision-making process of divide and conquer is implemented. A random forest is made up of many decision trees. Each tree is trained by using different categories of data in the training data set. The output is determined based on the results of all decision tree voting. Decision tree has obvious advantages for individual learning, it belongs to weak learners and which generalization ability is not strong. Random forest is composed of multiple learners with a good generalization of integrated learning, so it often appears in research together with SVM technology for comparison. Lerman [20] used RF for SCA and compared it with typical supervised learning methods and TAs to the representing neural networks and support vector machines. The result is that the RF comprehensive comparison score is highest when the amount of model data is small and the actual test data dimension is high, which proves that it is not always the best choice while the SVM is popular. In addition, Lerman [26] successfully used RF to break the masked AES, which also had obvious advantages when compared with TA.

\(k\)NN: Without any training, the data set is assigned a new instance to the class containing most \(k\) (fixed parameter) nearest neighbors by majority vote, which is supervised "lazy learning" [36]. Martinasek [35] used the \(k\)NN method for modeling power analysis and compared it to the classic TA. The experimental result shows that the \(k\)NN corresponding to the covariance matrix does not have a higher success rate, and the advantages are obvious. Compared with the TA, it can not select more feature points due to the restriction, bypassing the redundant stage of learning and training, and is a potential SCA in machine learning.

Verification. Verify the success rate of our training results by validating the data set and evaluate it based on the results. This step is usually the specific implementation phase of our evaluation of a machine learning model.

(5) Test matching. After the modeling phase, the training-verified model is matched with the measured attack data for analysis.

It is obvious that the supervised learning mechanism fits perfectly with the profiling SCA, so most of the whole machine-based profiling SCA fields use the generalized SVM and the random-featured RF model for training classification. Compared to the dimensional explosion problem of traditional modeling classes, the machine learning method is very good, and the data can be directly trained without presupposing the model. But when the data dimension is kept at a reasonably low level, traditional modeling attacks still dominate, which is related to the training time of machine learning.

4. Summary

With the rapid development of technology, people pay more and more attention to the field of encryption security. As the most powerful profiling SCA in theory, its development has been attracting in the industry.

The traditional classic profiling SCA process has been introduced in detail. The Gaussian profiling SCA is the most widely used and has been regarded as the representative of profiling SCA. In view of the fatigue of high-dimensional data processing by the classic Gaussian SCA, the advantages of machine learning are highlighted. Many model algorithms have been developed and applied in this field, and significant optimization results have been achieved. At present, there are still many shortcomings in the research of machine-based profiling SCA. The following are summarized:
(1) The supervised learning model is applied in a single way. The research history of machine learning modeling attacks can also be called the research history of SVM and RF. It is a very potential research direction for other deep neural networks and lazy learning (kNN and Bayesian classifier).

(2) The overall development of machine learning is not comprehensive. The correspondence between the nature of modeling and supervised learning has led to a high proportion of supervised learning. We can still use the semi-supervised learning and reinforcement learning with this nature to develop new methods.

(3) Insufficient anti-protection attacks. Most of the machine learning-based profiling attacks target unprotected encryption algorithms in order to compare them with traditional modeling attacks. However, there are many encryption devices with protection in the actual situation, and the comprehensive evaluation of a class of machine learning models should also add anti-protection effects.

Acknowledgments
I would like to show my deepest gratitude to my supervisor, Pro. Chen, a respectable and responsible scholar I have never met. Also it really gave me a lot of favors from my colleague Jiawei Ma. Without them, I can not achieve the ending.

References
[1] P. C. Kocher, J. Jaffe, and B. Jun, “Differential Power Analysis,” [C] in CRYPTO, ser. Lecture Notes in Computer Science, M. J. Wiener, Ed., vol. 1666. Springer, 1999, pp. 388–397.
[2] S. Chari, J. R. Rao, and P. Rohatgi, “Template Attacks,” [C] in CHES, ser. Lecture Notes in Computer Science, B. S. K. Jr., Cetin Kaya Koc?, and C. Paar, Eds., vol. 2523. Springer, 2002, pp. 13–28.
[3] W. Schindler, K. Lemke, and C. Paar, “A Stochastic Model for Differential Side Channel Cryptanalysis,” [C] in CHES, ser. Lecture Notes in Computer Science, J. R. Rao and B. Sunar, Eds., vol. 3659. Springer, 2005, pp. 30–46.
[4] E. Brier, C. Clavier, and F. Olivier, “Correlation Power Analysis with a Leakage Model,” [C] in CHES, ser. Lecture Notes in Computer Science, M. Joye and J.-J. Quisquater, Eds., vol. 3156. Springer, 2004, pp. 16–29.
[5] Kai Schramm, Gregor Leander, Patrick Felke, and Christof Paar. A Collision-Attack on AES Combining Side Channel- and Differential-Attack[C] in CHES 2004, LNCS 3156, pp. 163–175, 2004.
[6] B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel, “Mutual Information Analysis,” [C] in CHES, ser. Lecture Notes in Computer Science, E. Oswald and P. Rohatgi, Eds., vol. 5154. Springer, 2008, pp. 426–442.
[7] C. Rechberger and E. Oswald. “Practical Template Attacks” [C]. In Workshop on Information Security Applications — WISA 2004, August 2004.
[8] Gierlichs B, Lemkerst K, Paar C. Templates vs. stochastic methods. A performance analysis for side channel cryptanalysis. [J]. Lecture Notes in Computer Science, 2008.
[9] Omar Choudary and Markus G. Kuhn. “Efficient Template Attacks” [C] in CARDIS Smart Card Research and Advanced Application Conference 2013, LNCS 8419, pp. 253–270, 2014. Springer.
[10] Fisher, R.A.: The statistical utilization of multiple measurements. Ann. Eugen. 8,376–386 (1938)
[11] Mitchell, T.M.: Machine Learning. McGraw-Hill, New York, USA (1997).
[12] P. Dayan, M. Sahani, and G. Deback, “Unsupervised Learning,” in The MIT Encyclopedia of the Cognitive Sciences. The MIT Press, 1999.
[13] X. Zhu, “Semi-Supervised Learning,” in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds. Springer, 2010, pp. 892–897.
[14] S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” Informatica (Slovenia), vol. 31, no. 3, pp. 249–268, 2007.
[15] Dhillon, I. S., Y. Guan, and B. Kulis. (2004). “Kernel k-means, Spectral Clustering and Normalized
Cuts” In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD), 226-231, Portland, OR.

[16] Jolliffe, I. T. Principal Component Analysis. Springer-Verlag,[M] New York, 1986.

[17] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. [M]The handbook of brain theory and neural networks, 3361(10):1995.

[18] L. Breiman, (2001), “Random Forests”, Machine Learning, volume 45 n.1, pages 5-32.

[19] Cortes and V. Vapnik. Support-vector networks. Mach. Learn., 20(3):273–297,Sept. 1995.49. B. Scholkopf and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. [J]MIT Press, Cambridge, MA, USA, 2001

[20] Aha D W . Lazy Learning[M]. Springer Netherlands, 1997.

[21] B. Schölkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[22] Hospodar G, Gierlichs B, Mulder E D, et al. Machine learning in side-channel analysis: a first study[J]. Journal of Cryptographic Engineering, 2011, 1(4):293.

[23] Heuser and M. Zohner. Intelligent Machine Homicide - Breaking Cryptographic Devices Using Support Vector Machines. [C]In W. Schindler and S. A. Huss, editors, COSADE, volume 7275 of LNCS, pages 249-264. Springer, 2012.

[24] Bartkewitz and K. Lemke-Rust, Efficient Template Attacks Based on Probabilistic Multi-Class Support Vector Machines, in CARDIS,ser. Lecture Notes in Computer Science, S. Mangard, Ed., vol. 7771. Springer, 2012, pp. 263–276.

[25] Mladenic, D., Brank, J., Grobelnik, M., et al. Feature selection using support vector machines. In The 27th Annual International ACM SIGIR Conference (SIGIR 2004). 2004, p. 234–241. Lerman, L., Bontempi, G., Markowitch, O. Side Channel Attack: An Approach Based on Machine Learning, Constructive Side-Channel Analysis and Secure Design[C] - COSADE 2011, CASED, Darmstadt 2011

[26] Lerman, L., Medeiros, S. F., Bontempi, G., et al. A machine learning approach against a masked AES. In Proceedings of the 12th International Conference of Smart Card Research and Advanced Applications CARDIS. Berlin (Germany), 2013, p. 61–75.

[27] HE, H., JAFFE, J., ZOU, L. Side channel crypt analysis using machine learning. Stanford, 2012.

[28] Kim, H.M., Kang, D.J., Kim, T.H.: Flexible key distribution for scada network using multi-agent system. Bio-inspired, Learning, and Intelligent Systems for Security, ECSI S Symposium on (2007) 29–34

[29] Lian, S., Sun, J., Wang, Z.: One-way hash function based on neural network. CoRR abs/0707.4032 (2007)

[30] Wang, Y.h., Shen, Z.d., Zhang, H.g., Pseudo random number generator based on hopfield neural network. (August 2006) 2810–2813

[31] Liu, N., Guo, D.: Security analysis of public-key encryption scheme based on neural networks and its implementing. In: Computational Intelligence and Security, 2006 International Conference on. Volume 2. (nov. 2006) 1327 –1330

[32] Mislovaty, R., Perchenok, Y., Kanter, I., Kinzel, W., Secure key-exchange protocol with an absence of injective functions. Phys. Rev. E 66 (Dec 2002) 066102

[33] Martinasek, Z., Zeman, V. Innovative method of the power analysis. Radio engineering, 2013, vol. 22, no. 2, p. 586–594.ISSN: 1210-2512

[34] Martinasek, Z., Hajny, J., Malina, L. Optimization of power analysis using neural network. In Proceedings of the 12th International Conference Smart Card Research and Advanced Applications CARDIS. Berlin (Germany), 2013, p. 94–107.

[35] Patel H, Baldwin R O. Random Forest profiling attack on advanced encryption standard[J]. International Journal of Applied Cryptography, 2014, 3(2):ied Cryptography.

[36] Martinasek Z, Zeman V, Malina L., et al. k-Nearest Neighbors Algorithm in Profiling Power Analysis Attacks[J]. Radioengineering, 2016, 25(2):365-382.