Research on the Method of Selecting the Optimal Feature Subset in Big Data for Energy Analysis Attack

Xiaoyi Duan, You Li, Chengyuan Liu, Xiuying Li(✉), Wenfeng Liu, and Guoqian Li

Beijing Electronic Science and Technology Institute, Beijing, China
lixiuying@besti.edu.cn

Abstract. At present, the application of machine learning in energy analysis attack is a research hot spot of energy analysis attack, and the selection of feature points is an important factor that affects the machine learning model, how to choose the optimal feature subset is a key factor related to the success or failure of energy analysis attack. AES algorithm emphasizes to increase the complexity of encrypted data by a large number of encryption rounds. It generally runs ten rounds of encryption operation, but the energy information studied by attackers is only a part of one round in ten rounds. Therefore, it is of great significance to effectively select the optimal feature subset with the least amount of data from a large number of data for energy analysis attacks. According to the characteristics of high-dimensional small features of energy data, this paper proposes a new optimal feature subset selection method-secondary feature selection method named F-RFECV. Firstly, the F-test is used to quickly eliminate a large number of irrelevant and redundant features to initially select candidate energy feature subsets, and then the redundant features are further eliminated by recursive feature elimination and cross validation, so as to obtain the optimal energy feature subset, which effectively realizes the problem of small feature recognition in high-dimensional features, thus improving the success rate of model attack in subsequent machine learning. Experiments show that the attack success rate can be increased by 17% by using the secondary feature selection method (F-RFECV).

Keywords: F-test · Recursive feature elimination · Selection of feature points · Energy analysis attack · Machine learning

1 Introduction

1.1 Relevant Work

In order to reduce the number of samples and the size of the template, some special points need to be selected as the feature points in the power trace. In the energy analysis attack, the common feature point selection principle is to select the point which contains the most information about the key-related operations as the feature point. In papers

P. Gladyshev et al. (Eds.): ICDF2C 2021, LNICST 441, pp. 109–126, 2022.
https://doi.org/10.1007/978-3-031-06365-7_7

[1–5], a generally accepted criterion is proposed for the feature point selection method of this principle, each clock cycle can only select one point as the feature point, because more points in the same clock cycle can not provide more information. If this generally accepted criterion is not followed, the classification performance of energy analysis attack will be poor. So far, many different methods of feature point selection have been introduced.

In 1996, Kocher et al. introduced the feature point selection method based on mutual information analysis [6] (MIA). This method is to evaluate the correlation between qualitative independent variables and qualitative dependent variables, and then select feature points; In 2001, Gandolfi et al. introduced the method based on Kolmogorov Smirnov Analysis [7] (KSA), which is a nonparametric test method. It quantifies the distance between the empirical cumulative distribution function (CDF) of two random variables to determine their similarity; In 2003, the method based on mean difference [8] (DOM) was introduced, because only the operation dependent component and data dependent component will affect the mean value. Therefore, the point where the mean value changes greatly is the key related operation point, that is, the feature point; In 2006, gierlichs et al. proposed a method based on the sum of squares of paired t-differences [4] (SOST). SOST is a statistic based on t-test, which is essentially the accumulation of t-values, that is, the accumulation of correlation degree; Archambeau et al. introduced the method based on principal component analysis [2] (PCA), the feature points are selected by projecting the data sets of multiple related features onto the coordinate system with fewer related features; In 2007, Mangard et al. introduced the method based on correlation power analysis (CPA) and the method based on signal-to-noise ratio (SNR) [9].

CPA is a method that uses Pearson correlation coefficient to help understand the relationship between features and response variables. This method measures the linear correlation between variables. SNR quantifies the amount of information leaked by a single point in the energy trace. The higher the SNR is, the easier it is to identify the available components from the noise, and the more information the point divulges. Therefore, high SNR points are selected as feature points; In 2008, Standaert et al. introduced the method based on Fisher linear discriminant analysis [10] (LDA). It is a supervised algorithm, similar to PCA, LDA also extracts a new coordinate axis, projects the original high-dimensional data to the low dimensional space, and directly optimizes the low dimensional space to obtain the best classification separability. In 2013, Mather et al. Proposed the variance based method [11] (VAR); Only the operation dependent component and the data dependent component will affect the difference. Therefore, the point with large variance is the key related operation point, that is, the feature point. In these methods, the point with the strongest signal strength estimation is selected as the feature point in each clock cycle. In 2014, Fan G proposed the method based on a select sample distribution similar to normal distribution points as the feature points of feature point selection method, the experimental results show that compared with the previous method, the template attack based on the feature points selected by this method has better classification performance [12]. In 2015, Roy proposed class-specific feature selection based on the same concept as LDA and MMC to maximize the distance between classes and minimize the distance between inner points of classes. The data were trained on the

distributed platform Apache Spark, and the efficiency of the proposed method for the selection of high-dimensional data feature points was proved [13].

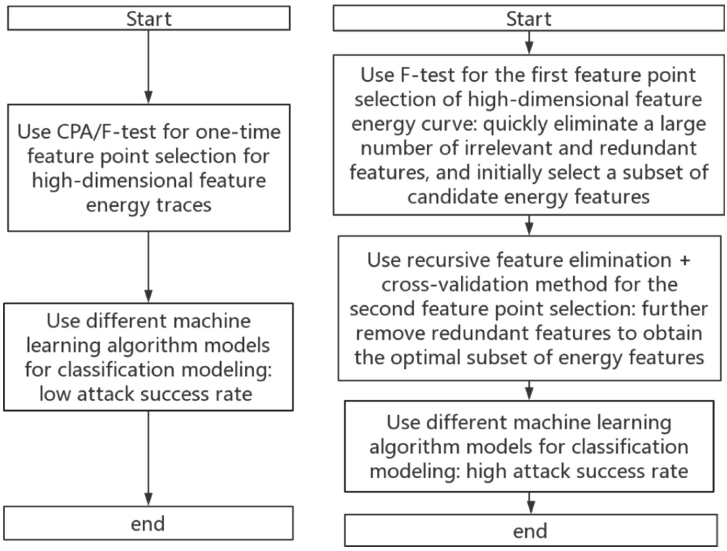
Research in 2017, He and others studied the feature selection of multivariate response coefficient models with a large number of covariates, aiming at the relatively limited data samples, it is difficult to estimate a large number of coefficient functions nonparametric, put forward a punishment least-squares framework, when only the relevant variables included in the model, the proposed method can uniformly identify relevant covariate, and the corresponding coefficient of convergence rate can be estimated with the same function [14]. In 2019, Ireneusz et al. used stack technology to improve the generalization ability of machine learning in processing feature point selection data in view of the imbalance of categories among data [15]. In 2020, Khosla et al. introduced topologically preserved distance scaling (TPDS) to enhance feature point selection, which aims to reproduce distance information in higher dimensions. In addition to providing better visualization than typical distance preservation methods, TPDS can also better classify data points in terms of narrowing feature points [16]. In addition, the paper [1] introduces that the energy analysis attack based on principal component analysis is inefficient due to its high computing requirements. The paper [11] introduces that the energy analysis attack based on principal component analysis may not improve the classification performance. The paper [17] introduce the rare condition that LDA based energy analysis attacks depend on equal covariance, which is not true for most encryption devices. Therefore, compared with PCA based energy analysis attack, it is not a better choice [17].

1.2 Our Contribution

At present, the existing feature selection methods are still facing difficulties in the field of high-dimensional small sample, and their full potential in energy analysis attack has not been fully utilized. In order to solve this problem, this paper proposes a new energy analysis attack feature point selection method. Compared with the previous methods, this method has a completely different basic principle. This paper proposes a feature point selection method (F-RFECV) based on F-test and recursive feature elimination + cross validation (RFECV), which combines the different advantages of filtering and wrappering feature selection methods to eliminate irrelevant and redundant features.

1. Firstly, F-test is used to evaluate correlation between feature points and class labels, quickly remove a lot of irrelevant and redundant features to initially select a subset of candidate energy features.
2. Then the recursive feature elimination + cross validation (RFECV) method is used to further eliminate redundant features and obtain the optimal subset of energy features, which effectively realizes the problem of small feature recognition in large data.

In addition, experiments show that our new feature point selection method has better classification performance than previous methods. Compared with the previous feature subset method, the proposed F-RFECV method can improve the attack success rate by 17%. Figure 1 compares the traditional feature point selection method with the F-RFECV feature point selection method proposed in this paper.



(a) Traditional feature point selection method

(b) Feature point selection of the F-RFECV method proposed in this paper

Fig. 1. Comparison of feature point selection methods

1.3 Structure of This Paper

The structure of this paper is as follows. In Sect. 2, the selection of attack point is introduced. In Sect. 3, this paper introduces the optimal feature subset selection methods in a large number of data, including feature point selection strategy, feature point selection based on F-test, recursive feature elimination + cross validation (RFECV) principle. In Sect. 4, this paper carries out experimental verification and discussion. In Sect. 5, this paper summarizes and looks forward to the future research.

2 Selection of Attack Points

In energy analysis attack, different attack points have different attack effects. Because the SBOX has the highest nonlinearity in various encryption algorithms, for energy analysis attacks, choosing SBOX as the attack point will get more information related to the algorithm. Therefore, in energy analysis attacks, SBOX output value is often selected as the attack point. When SBOX output is selected as the attack point, it can be divided into SBOX output Hamming weight model and SBOX output bit by bit model. Different attack models have different attack effects.

2.1 SBOX Output Hamming Weight Model

Hamming weight refers to the number of non-zero elements in a string. For the commonly used binary string, it is the number of 1 in the string. In energy analysis attacks,

Hamming weight model is generally used to represent the power consumption model of the operation data in the chip. Establish the label of energy trace and SBOX output Hamming weight, the key of cryptographic chip is further obtained by obtaining the Hamming weight of SBOX output. Taking AES algorithm as an example, for each byte of plaintext input, the output value of SBOX is in the interval of [0, 255], and they are not equal to each other. Hamming weight refers to the number of 1 in a binary value, so there are 9 kinds of Hamming weight output by SBOX.

2.2 SBOX Output Bit-by-Bit Attack Model

When the Hamming weight model is used to attack, it is necessary to first train a 9-classification model according to the Hamming weight label. Compared with the Hamming weight model, the Bit-by-bit model only needs binary classification per bit, and the number of classifications is less. Therefore, it is easier to classify by machine learning and can obtain better classification accuracy.

At the same time, when the Hamming weight model is used for the attack, the key cannot be obtained directly by obtaining the intermediate Hamming weight, so further analysis is needed to obtain the key. The common method is to carry out enumeration attack again, so it requires multiple energy traces to attack successfully. Compared with the Hamming weight model, the Bit-by-bit model needs only one energy trace to successfully attack, so the bit-by-bit model has higher efficiency and lower cost.

In conclusion, compared with the Hamming weight model, the Bit-by-bit model has better classification accuracy, higher efficiency and lower cost, so this paper chooses the Bit-by-bit model to attack. The comparison between the Hamming weight model and the bit-by-bit attack model is shown in Table 1.

Table 1. Comparison between Hamming weight model and bit-by-bit attack model

	The correct rate	The efficiency	The cost
Hamming weight model	Low	Low	High
Bit-by-bit attack model	High	High	Low

3 The Optimal Feature Subset Selection Method in Big Data

3.1 Feature Point Selection Strategy

Feature selection is a search optimization problem. For a data set of N features, the feature selection strategy is to find the optimal feature subset in 2N-1 non-empty feature subsets. The search strategy can be roughly divided into three categories:

(1) Exhaustive search. Exhaustive search refers to searching the feature subsets of all feature space sets. Its advantage is clear, that is, an optimal subset can be obtained, but this is limited to a data set with a small number of feature points; When the number of features of data sets is large, this method will bring disasters, that is, huge computing overhead and huge time cost.

(2) Heuristic search. Heuristic search avoids simple and brutal exhaustive search, but continuously adds or deletes features to the current feature subset according to a specific order in the search process, and finally obtains the best feature subset.
(3) Random search. Random search starts with a certain candidate feature subset generated randomly, and gradually approaches the global optimal solution according to certain heuristic information and rules.

The method of feature selection can be divided into three categories according to whether it is independent of the subsequent learning algorithm:

(1) Filter: The idea is to evaluate the pros and cons of each feature in the data set based on the data characteristics of the data set based on statistics, information theory and other disciplines, and finally sort the pros and cons of each feature, and select several better features to form the final feature subset. This paper can see that the filtering method relies on the data characteristics of the data set, rather than using specific learning algorithms to evaluate feature subsets.
(2) Wrapper: The idea of Wrapper method is to establish a machine learning model, and evaluate the pros and cons of this subset based on the prediction performance of the test feature subset in the machine learning model, and select or delete some features according to the pros and cons of the prediction performance. This method focuses on the pros and cons of the entire feature subset, and does not focus on the pros and cons of each feature in the feature subset. Therefore, the optimal subset obtained by this method does not mean that every feature in itself is optimal.
(3) Embedded: The idea of the embedded method is to first establish a machine learning model and train it, then obtain the weight coefficient of each feature and sort it, and finally select the feature point with a larger weight coefficient as the optimal subset.

This paper combines the different advantages of filtering and wrappering to eliminate irrelevant and redundant features, and proposes an energy feature selection method based on F-test and recursive feature elimination + cross validation method (F-RFECV). Firstly, the F-test is used to quickly eliminate a large number of irrelevant and redundant features to initially select candidate energy feature subsets, and then the redundant features are further eliminated by recursive feature elimination + cross validation (RFECV), so as to obtain the optimal energy feature subset, which effectively realizes the problem of small feature recognition in high-dimensional features.

3.2 First Feature Point Selection Based on F-test

The filtering method obtains the evaluation criteria of feature selection by using the intrinsic association of the data set itself, which is independent of the specific learning algorithm, and has good universality and low algorithm complexity. As a feature prefilter, it is suitable for quickly eliminating a large number of irrelevant features from large-scale data sets. For big data, it is suitable for fast feature selection to select candidate feature points. In this paper, F-test based on filtering method is used to preliminarily select feature subset.

F-test, also known as joint hypothesis test, was proposed by Fisher, a British statistician, which is a test method to test whether the total variance of two normal random variables are equal. This paper uses the evaluation parameter F-score of F-test to select features. If two random variables $X = \{x_1, x_2, x_3, ..., x_n\}$ and $Y = \{y_1, y_2, y_3, ..., y_n\}$ obey normal distribution, the mean values of these two sequences can be obtained as follows:

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1}$$

$$\overline{\mathrm{Y}}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{2}$$

Therefore, it can be obtained the variances of the two sequences as follows:

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \tag{3}$$

$$S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \overline{Y})^2 \tag{4}$$

At last, it can be obtained the formula of F(n−1, m−1) distribution as follows:

$$F = \frac{S_X^2}{S_Y^2} \tag{5}$$

For the data selected in the next section of this paper, each energy trace contains 435002 points, so it is suitable for fast feature point primary selection with F-test.

In this section, the feature points of the first bit and the second bit of SBOX output value are selected by F-test. The experimental results are shown in Fig. 2 and Fig. 3.

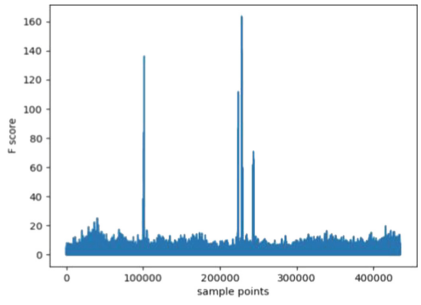


Fig. 2. F-test score of the first bit of Sbox output value in the first round

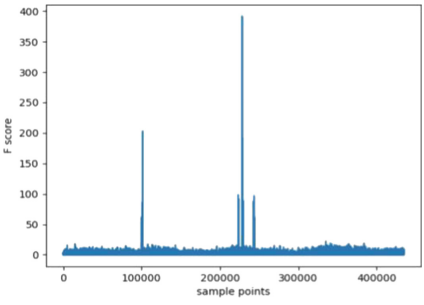


Fig. 3. F-test score of the second bit of Sbox output value in the first round

From Fig. 2 and Fig. 3, we can see that when the sampling point is performing key correlation calculation, there will be a spike at the sampling point. We select the sampling point as the feature point. We observe carefully and find that the feature points selected by the two bits are different, so we should select their corresponding features separately for each bit.

3.3 Second Feature Point Selection Based on Recursive Feature Elimination and Cross Validation

After the first feature point selection based on filtering, the feature subset can be initially selected. However, because the threshold value of the feature point method based on filtering is not easy to select, the feature points selected are still more, and it is not easy to effectively select the best feature points. Therefore, this article uses the characteristic advantages of the wrappering method mentioned above to perform the second feature selection on the basis of the selected feature subset, so as to achieve the purpose of selecting the optimal feature subset.

Recursive feature elimination (RFE) is a kind of wrappering feature selection algorithm. The principle of recursive feature elimination is to use machine learning algorithm to build a model and carry out multiple trainings. After each training, the features with smaller weights are eliminated by comparing the weight coefficients of each feature point. The above operations are performed on the new data set until all features are traversed, and the final subset is the final feature selection result.

In this paper, the classical SVM-RFE algorithm is used to select the best feature points, so this algorithm is used to describe the process of feature selection. First of all, in each round of training, all features N will be selected for training, so we get N classified hyperplanes $w * x + b = 0$, After the first training, the number of feature points with the smallest square value of the component in the weight w is deleted by using the SVM-RFE algorithm, then getting a data set with N-1 feature points; In the second training, the $N-1$ dimensional data set is input for training, and the algorithm will continue to remove the feature corresponding to the smallest square value of the middle component in the weight w. And so on, until the remaining feature numbers meet our requirements. The above optimal feature points seem to be perfect? Of course not. This paper realizes that before selecting RFE feature points, fixing the selected feature points is needed, but the best feature points are unknown. When this paper executes SVM-RFE through cross validation, the problem will be solved. Cross validation means that when the machine learning algorithm produces over-fitting phenomenon, the data sets are grouped, with a large part as the training set and a small part as the verification set. Firstly, the machine learning algorithm is trained by the training set, and then the trained model is verified by the verification set, and the parameters such as accuracy rate are calculated to evaluate the performance of machine learning classification. After the SVM-RFE algorithm is cross verified, this paper can get the parameter validation error of all subsets calculated by the algorithm, and the subset with the smallest validation error is the best classification subset. As shown in Fig. 4.

The abscissa of the above figure is the feature points of the data, and the ordinate is the cross validation score of the corresponding feature point subset. It can be seen from the figure that the best feature point selected by RFECV algorithm is 39.

4 Experimental Results and Analysis

The experimental data in this paper comes from the fourth stage of the international academic Contest of differential power analysis (DPA Contest V4, DPA_V4), which is collected from the AES-256 cryptography algorithm running on the ATmega163 chip,

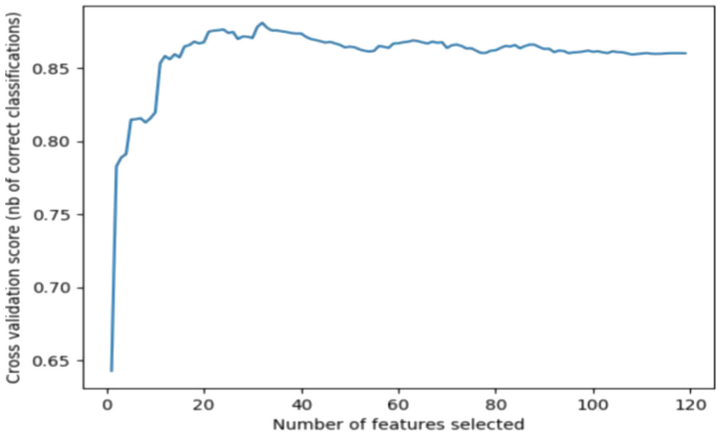


Fig. 4. Selection of feature points of RFECV

each trace contains 435002 points, In this paper, a total of 1000 energy traces in the data set are used for experiments. The experimental attack model is SBOX output Bit-by-bit model.

In order to verify the advantages of the proposed method, this paper compares the two feature point selection methods proposed with other methods. In this paper, Four machine learning algorithms, Support Vector Machine (SVC), Linear Discriminant Analysis (LDA), Decision Tree Classifier (DTC) and Logistic Regression Classifier (LRC) are used to build classification models on the candidate and optimal feature subsets, and the classification accuracy is used to evaluate the performance of different methods. The main experiments are as follows:

1) Filter method is used to select the first feature points to get the candidate feature subset. In order to verify the advantages of F-test based on filtering method, this paper compares it with Pearson correlation coefficient (PCC).
2) Secondly, on the basis of the candidate feature subset, this paper uses the wrappering method for the second feature point selection to get the optimal feature subset. In order to verify the advantages of recursive feature elimination and cross validation (RFECV) based on wrappering method, this paper compares it with Principal component analysis (PCA).
3) This paper compares the secondary feature selection method F-RFECV proposed in this paper with the primary feature selection method F-test.

4.1 Comparison of Feature Selection Between F-test and PCC Method

PCC and F-test are based on the mathematical features of data and label points, and both methods can distinguish target features from noise features. In this paper, the above two methods are used to select feature points, and four machine learning algorithms, SVC, LDA, DTC and LRC are used to classify and model the data after feature point selection, so as to observe which feature point selection method is more conducive to the energy analysis attack of the data set.

It can be seen from Sect. 3.2 that the feature points selected for each bit of SBOX output are different, so we select the feature points for each bit of SBOX output separately. Because there are 435002 points on each energy curve of the data selected in this experiment, the amount of data is too large. Therefore, based on the feature point selection of F-test, we only select the points whose evaluation parameter F-score value is greater than 50; Based on PCC, we only select the feature points whose absolute value of evaluation coefficient is greater than 0.2; The experimental results are shown in Table 2 and Fig. 5 below.

Table 2. Comparison of feature points selected by PCC and F-test

Bit	1	2	3	4	5	6	7	8
The dimensions chosen by PCC	148	228	154	252	239	156	265	278
The dimensions chosen by F-test	130	176	117	212	182	119	222	240

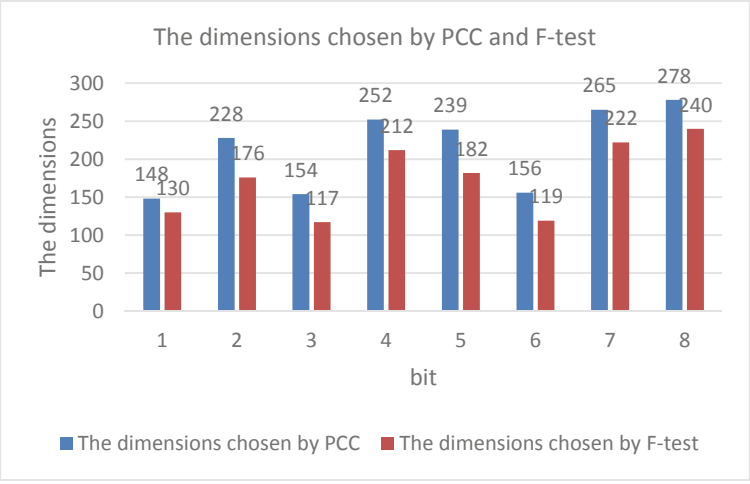


Fig. 5. The dimensions chosen by PCC and F-test

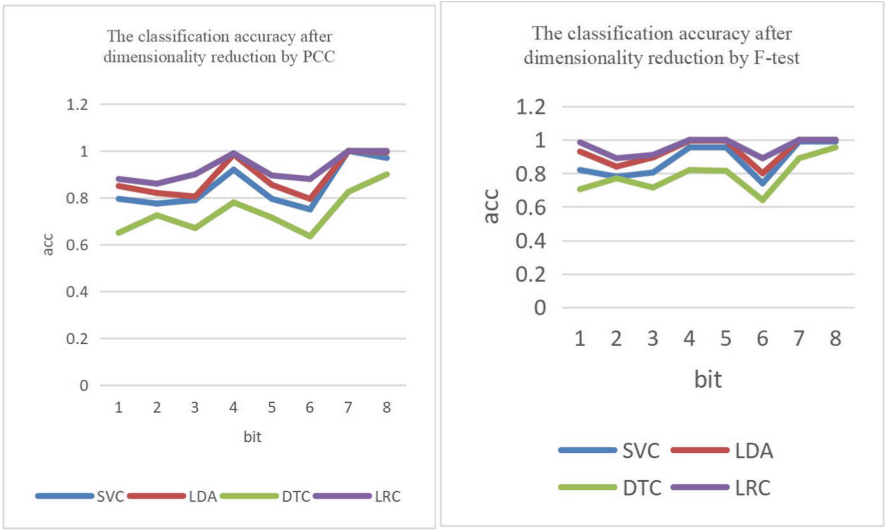
From Table 2 and Fig. 5, we can find that on the basis of distinguishing target features and noise features, the two feature point selection methods retain the mathematical characteristics of the data to the maximum extent, and the number of feature points has been greatly reduced. When F-test is used to select feature points, the feature points of each bit are smaller than those of PCC. However, does less feature points mean that it is a better feature subset? The answer is negative. Therefore, we further verified it.

In this paper, four machine learning algorithms, SVC, LDA, DTC and LRC, are used to classify and model the feature point selection results after PCC and F-test, and the classification accuracy is shown in the following Table 3 and Fig. 6:

From Table 3 and Fig. 6, we can see that although the feature points obtained by F-test feature point selection method are smaller than those obtained by PCC feature

Table 3. Classification of PCC and F-test feature points after selection

Number of bits	PCC				F-test			
	SVC	LDA	DTC	LRC	SVC	LDA	DTC	LRC
1	0.795	0.85	0.65	0.88	0.82	0.93	0.705	0.985
2	0.775	0.82	0.725	0.86	0.78	0.84	0.77	0.89
3	0.79	0.805	0.67	0.9	0.805	0.895	0.715	0.91
4	0.92	0.985	0.78	0.99	0.955	0.995	0.82	1
5	0.795	0.855	0.715	0.895	0.955	0.995	0.815	1
6	0.75	0.795	0.635	0.88	0.74	0.8	0.64	0.89
7	1	1	0.825	1	0.99	1	0.89	1
8	0.97	0.995	0.9	1	0.99	1	0.955	1



(a) classification accuracy after PCC feature point selection

(b) classification accuracy after F-test feature point selection

Fig. 6. Classification of PCC and F-test feature points after selection

point selection method, after selecting feature points by this method, the classification accuracy is still higher than PCC. Therefore, through the above experiments, this paper gets a conclusion: F-test feature point selection method is superior to PCC feature point selection method for single-bit preliminary feature point selection.

However, we further founds that a large number of irrelevant and redundant features were quickly eliminated by using F-test, and the feature-related points were initially selected. Although its feature point selection method is superior to PCC feature point selection method, its overall accuracy rate is not high. Therefore, this paper carries out

the second feature point selection to further eliminate redundant features in order to obtain the optimal energy feature subset.

4.2 Comparison of Feature Selection Between Recursive Feature Elimination Plus Cross Validation (RFECV) and Principal Component Analysis (PCA)

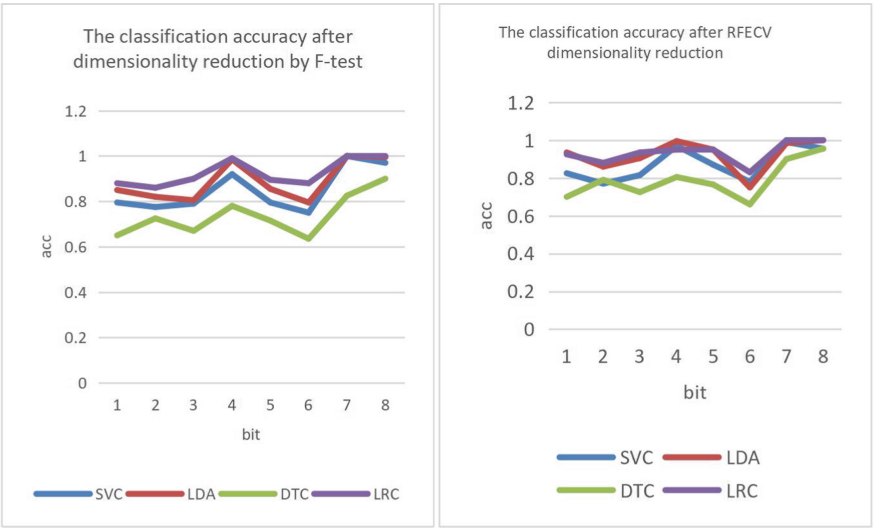
After the first step of feature point selection, the classification result of the subset this paper obtained seems to be acceptable, but it cannot achieve the optimal classification result. Therefore, this paper introduces the second step of feature point selection, In this paper, the wrappering method is used for the second feature point selection to obtain the optimal feature subset. In order to verify the advantages of the RFECV based on the wrappering method proposed in this paper, this paper compares it with the PCA method.

In Sect. 4.1, on the basis of distinguishing the target features and noise features, through the F-test one-time feature point selection method, the mathematical features of the data are retained to the maximum extent, and an optimal feature subset is obtained, as shown in Table 2. But is the subset of the best candidate feature points obtained in the first F-test the best input for the second feature point selection RFECV? The answer is unknown. In this regard, we conducted futher experiments. The experimental results are shown in Table 4 and Fig. 7, which compares the classification accuracy after one F-test feature point selection and two feature points selection.

Table 4. Comparison of F-test feature point selection and RFECV feature point selection classification

Bit	F-test				RFECV			
	SVC	LDA	DTC	LRC	SVC	LDA	DTC	LRC
1	0.82	0.93	0.705	0.985	0.825	0.935	0.7	0.925
2	0.78	0.84	0.77	0.89	0.77	0.86	0.79	0.88
3	0.805	0.895	0.715	0.91	0.815	0.905	0.725	0.935
4	0.955	0.995	0.82	1	0.97	0.995	0.805	0.95
5	0.955	0.995	0.815	1	0.87	0.95	0.765	0.95
6	0.74	0.8	0.64	0.89	0.78	0.75	0.66	0.83
7	0.99	1	0.89	1	0.99	0.985	0.9	1
8	0.99	1	0.955	1	0.955	1	0.955	1

From Table 4 and Fig. 7, we find that after the first F-test feature point selection, the second feature point selection makes the classification accuracy of the machine learning algorithm seem to be lower. Then what went wrong? Obviously the first step of feature point selection is the problem. That is, for the second feature point selection method RFECV proposed in this paper, the best candidate feature point subset obtained by the first F-test feature point selection is not the best input for the second feature point selection.



(a) Classification accuracy after feature point selection by F-test

(b) Classification accuracy after feature point selection of RFECV

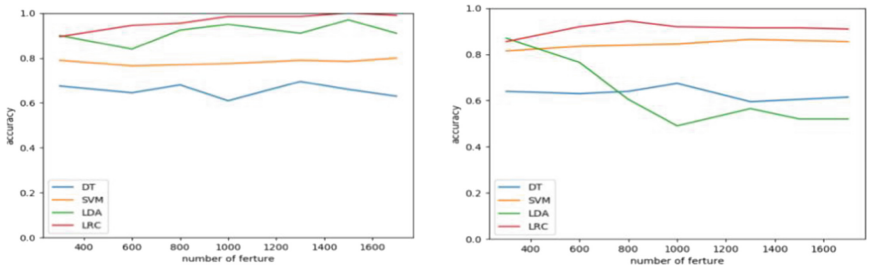
Fig. 7. Comparison of classification accuracy after feature point selection by F-test and RFECV

In order to obtain the best input for the second feature point selection, we use F-test to obtain different feature subsets for experiments. As shown in Fig. 7, we find that the classification accuracy of the feature point selection of the 6th bit twice is the smallest. Therefore, we use F-test to obtain different feature subsets for the 6th bit, so as to obtain the best input for the second feature point selection. However, because the feature point selection based on RFECV takes too long, we only obtain the feature subset of 300, 600, 800, 1000, 1300, 1500 and 1700 feature points as the input of the second feature point selection. The PCA feature point selection method is compared with the work.

Four machine learning algorithm models of SVC, LDA, DTC and LRC are used for classification, and the experimental results are shown in Fig. 8. The classification accuracy of the second feature point selection using RFECV method is shown in Fig. 8(a). The classification accuracy of PCA method for the second feature point selection is shown in Fig. 8(b).

From Fig. 8, we can draw the following conclusions:

1) The different feature subsets obtained by the first F-test feature point selection have different effects on the secondary feature point selection.
2) Comparing with Fig. 8(a) and Fig. 8(b), we can see that when using DT, LDA and LRC machine learning algorithms, the classification effect of feature set based on RFECV is better than PCA, while for SVM classification, the classification effect of feature set based on RFECV feature point selection is slightly lower than that of PCA. Generally speaking, RFECV is superior to PCA in feature point selection.
3) According to Fig. 8(a), it can be seen that when the RFECV method is selected for the second feature point selection, with the increase of the number of features



(a) Classification accuracy after RFECV feature point selection

(b) Classification accuracy after PCA feature point selection

Fig. 8. Classification accuracy after secondary feature point selection with different number of feature points

Table 5. The classification accuracy of the first eight bits

Bit	Feature points											
	1000				1300				1500			
	DT	SVM	LDA	LRC	DT	SVM	LDA	LRC	DT	SVM	LDA	LRC
1	0.735	0.82	0.985	1	0.715	0.88	1	1	0.745	0.855	1	0.99
2	0.74	0.81	0.99	0.995	0.755	0.81	0.99	0.98	0.73	0.83	0.995	0.99
3	0.715	0.8	0.795	0.945	0.71	0.84	0.97	0.995	0.725	0.865	0.99	0.98
4	0.735	0.885	0.985	1	0.795	0.985	1	1	0.83	0.995	1	0.99
5	0.825	0.975	1	1	0.77	0.88	1	1	0.715	0.89	0.975	1
6	0.645	0.77	0.95	0.985	0.66	0.79	0.91	0.985	0.67	0.785	0.97	1
7	0.935	1	1	1	0.865	1	1	1	0.9	1	0.995	0.98
8	0.97	1	1	1	0.99	1	1	1	0.985	1	1	1

selected in the first feature point selection, the classification accuracy after the two feature points selection also increases. Among them, when the number of feature points selected in the first time is 1500, the classification accuracy can reach 100% when using LRC method to classify the results after two feature points selection. In general, when the feature number of the input feature subset selected for the second feature point is maintained at 1000 to 1500, the classification result is optimal.

In order to obtain the optimal feature subset of all 8-bit in the first feature point selection by F-test, we carried out further experiments. At this time, we only carried out experiments when the feature points were maintained between 1000 and 1500. The experimental results are shown in Table 5 and Fig. 9. Table 5 and Fig. 9 list the classification accuracy of the feature subset of the first eight bits after two feature point selection by selecting different feature points in the first feature point selection.

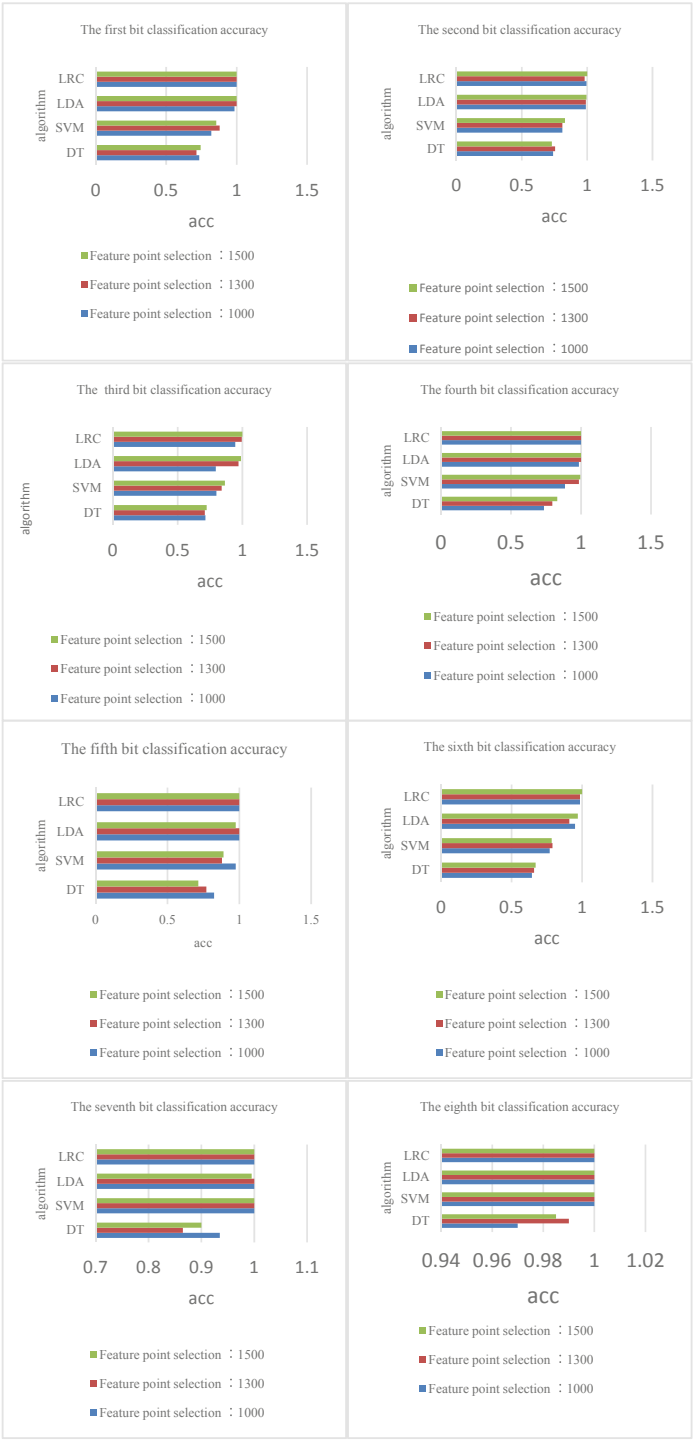


Fig. 9. Classification accuracy of the first eight bits

It can be seen from Table 5 and Fig. 9 that, in general, for each bit value of SBOX output, when the candidate energy feature subset initially selected by F-test contains 1500 feature points, the classification accuracy of different machine learning algorithms can reach the highest through RFECV, that is, the optimal energy feature subset can be obtained under this condition.

Finally, under the condition of optimal energy feature subset, this paper compares the proposed secondary feature point selection method F-RFECV with the primary feature point selection method F-test, and the experimental results are shown in Table 6.

Table 6. Comparison of classification accuracy between F-test primary feature point selection and F-RFECV secondary feature point selection proposed in this paper

Bit	Primary feature point selection and secondary feature point selection							
	F-test one-time feature point selection				F-RFECV secondary feature point selection			
	SVC	LDA	DTC	LRC	SVM	LDA	DTC	LRC
1	0.82	0.93	0.705	0.985	0.855	1	0.745	0.99
2	0.78	0.84	0.77	0.89	0.83	0.995	0.73	0.98
3	0.805	0.895	0.715	0.91	0.865	0.99	0.725	0.99
4	0.955	0.995	0.82	1	0.995	1	0.83	1
5	0.955	0.995	0.815	1	0.89	0.975	0.715	1
6	0.74	0.8	0.64	0.89	0.785	0.97	0.67	0.97
7	0.99	1	0.89	1	1	0.995	0.9	1
8	0.99	1	0.955	1	1	1	0.985	1

Table 6 compares the classification accuracy after F-test primary feature point selection and the F-RFECV secondary feature point selection proposed in this paper (the number of feature points selected is 1500). From Table 6, It can be seen that after the second feature point selection, the classification accuracy of the energy trace using the machine learning algorithm is improved. This paper calculates that the attack success rate of the F-RFECV proposed in this paper can be increased by 17%. Therefore, we believe that the two-time feature point selection of this scheme is superior to the scheme using only one feature point selection.

5 Summary

This paper determines the best model of feature point selection for cryptographic algorithm through two-time feature point selections, namely feature selection based on F-test and RFECV. F-test is used for the first feature point selection, 1500 feature points are selected as the input of the second feature point selection, and then the redundant features are further eliminated by RFECV, so as to obtain the optimal energy feature subset,

which effectively realizes the problem of small feature recognition in high-dimensional features, thus improving the success rate of model attack in subsequent machine learning. Experiments show that the attack success rate can be increased by 17% by using the secondary feature selection method (F-RFECV).

Acknowledgments. This research was supported by the High-tech discipline construction funds of China (No. 20210032Z0401), the High-tech discipline construction funds of China (No. 20210033Z0402) and the open project of Key Laboratory of cryptography and information security in Guangxi, China (No. GCIS201912).

References

1. Rechberger, C., Oswald, E.: Practical template attacks. In: Lim, C.H., Yung, M. (eds.) WISA 2004. LNCS, vol. 3325, pp. 440–456. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31815-6_35
2. Archambeau, C., Peeters, E., Standaert, F. -X., Quisquater, J. -J.: Template attacks in principal subspaces. In: Goubin, L., Matsui, M. (eds.) CHES 2006. LNCS, vol. 4249, pp. 1–14. Springer, Heidelberg (2006). https://doi.org/10.1007/11894063_1
3. B¨ar, M., Drexler, H., Pulkus, J.: Improved template attacks. In: COSADE2010 (2010)
4. Gierlichs, B., Lemke-Rust, K., Paar, C.: Templates vs. stochastic methods. In: Goubin, L., Matsui, M. (eds.) CHES 2006. LNCS, vol. 4249, pp. 15–29. Springer, Heidelberg (2006). https://doi.org/10.1007/11894063_2
5. Hanley, N., Tunstall, M., Marnane, W.P.: Unknown plaintext template attacks. In: Youm, H.Y., Yung, M. (eds.) WISA 2009. LNCS, vol. 5932, pp. 148–162. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10838-9_12
6. Kocher, P.C.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In: Koblitz, N. (ed.) CRYPTO 1996. LNCS, vol. 1109, pp. 104–113. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-68697-5_9
7. Gandolfi, K., Mourtel, C., Olivier, F.: Electromagnetic analysis: concrete results. In: Koç, Ç.K., Naccache, D., Paar, C. (eds.) CHES 2001. LNCS, vol. 2162, pp. 251–261. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44709-1_21
8. Chari, S., Rao, J.R., Rohatgi, P.: Template attacks. In: Kaliski, B.S., Koç, çK., Paar, C. (eds.) CHES 2002. LNCS, vol. 2523, pp. 13–28. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36400-5_3
9. Mangard, S., Oswald, E., Popp, T.: Power Analysis Attacks: Revealing the Secrets of Smart Cards. Springer, Berlin (2007).https://doi.org/10.1007/978-0-387-38162-6
10. Standaert, F.-X., Archambeau, C.: Using subspace-based template attacks to compare and combine power and electromagnetic information leakages. In: Oswald, E., Rohatgi, P. (eds.) CHES 2008. LNCS, vol. 5154, pp. 411–425. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85053-3_26
11. Montminy, D.P., Baldwin, R.O., Temple, M.A., Laspe, E.D.: Improving cross- device attacks using zero-mean unit-variance mormalization. J. Cryptographic Eng. **3**(2), 99–110 (2013)
12. Fan, G., Zhou, Y., Zhang, H., Feng, D.: How to choose interesting points for template attacks more effectively? In: Yung, M., Zhu, L., Yang, Y. (eds.) INTRUST 2014. LNCS, vol. 9473, pp. 168–183. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27998-5_11
13. Roy, A.: A classification algorithm for high-dimensional data. Procedia Comput. Sci. **53**, 345–355 (2015)

14. He, K., Lian, H., Ma, S.: Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. J. Am. Statist. Assoc. **113**, 746, 754 (2017)
15. Ireneusz, C., Piotr, J., Thanh, N.N., Edward, S., Bogdan, T., Van Du, N.: Data reduction and stacking for imbalanced data classification. J. Intell. Fuzzy Syst. **37**(6), 7239 (2019)
16. Khosla, K., Jha, I.P., Kumar, A., Kumar, V.: Local-topology-based scaling for distance preserving dimension reduction method to improve classification of biomedical data-sets. Algorithms **13**(8), 192 (2020). https://doi.org/10.3390/a13080192
17. Choudary, O., Kuhn, M.G.: Efficient template attacks. In: Francillon, A., Rohatgi, P. (eds.) CARDIS 2013. LNCS, vol. 8419, pp. 253–270. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08302-5_17