Distribution-based Adversarial Filter Feature Selection against Evasion Attack

Patrick P. K. Chan*, YuanChao Liang*, Fei Zhang†, Daniel S. Yeung‡
*School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
†College of Computer and Information Engineering, Henan Normal University, Xinxiang, China
‡Hong Kong
Email: patrickchan@ieee.org, 754571662@qq.com, zhangfei@htu.edu.cn, danyeung@ieee.org

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9533763

***Abstract*—Feature selection plays an important role in machine learning in order to reduce model complexity and extract more meaningful information. The recent studies indicate that not only the generalization ability but also the security should be considered in selecting features in an adversarial environment in which a classifier may be misled by an adversary intentionally. However, since only the nearest legitimate sample to a malicious sample is considered, the existing adversarial filter feature selection method is sensitive to outlier samples and is not suitable for Boolean features. A distribution-based adversarial filter feature selection method is proposed against evasion attack in this study. Our method uses distribution-based measurements, such as Symmetric Uncertainty and Earth Mover's Distance, are used to quantify the generalization ability and security of a feature subset. The experiments suggest our proposed method outperforms existing methods in terms of robustness and stability in datasets with Boolean and real-valued features.**

I. Introduction

Machine learning techniques have been applied to many applications although their security issues have not been investigated clearly. Many studies [1]–[4] indicate that machine learning, even for advanced techniques like Deep Learning, downgrades significantly in an adversarial environment in which an adversary misleads decisions of a system by manipulating samples. A sample with slight modification may also change the output of a system. For example, Convolutional Neural Network (CNN) can be fooled by an image with unrecognizable modification [5]. This is because an adversarial attack violates the assumption on the same (or similar) distribution of the training and testing samples of the traditional machine learning methods [6], [7].

Evasion attacks [1], [3], aiming to mislead a trained classifier by carefully crafting a sample in the test phase, is an adversarial attack commonly found in many applications [6], [8], [9]. For example, an automated driving system ignores a road sign with a specifically designed sticker [4], [10]. Most countermeasures against evasion attack focus on how to train a robust classifier [11]–[13] by optimizing not only the generalization ability but also the robustness under attacks [2], [14]. Another countermeasure is to filter the suspected samples before training [15]. Adversarial feature selection [16]–[19], which is an example of data sanitization, chooses the features with good discriminative ability and robustness against attacks.

Although wrapper approaches [16], [17] of the adversarial feature selection achieve satisfying performance, evaluating a feature candidate experimentally is time consuming, especially for the robustness term. An embedded adversarial feature selection method [18] optimizes the training parameters and feature subset simultaneously. However, this method is only designed for a linear classifier and may not be generalized to other kinds of classifiers easily. To the best of our knowledge, there is only one adversarial filter approach, *i.e.,* Adversarial filter feature selection method based on mRMR (FAFS) [19], against evasion attack. In FAFS, a feature candidate is evaluated by the generalization ability and security term. Feature redundancy and class relevance are captured by the generalization ability term, while the security term calculates the averaged distance from each malicious sample to its nearest legitimate sample. As a filter approach, the computational complexity is much lower than a wrapper approach since the evaluation does not rely on classifier training. However, FAFS is sensitive to outliers as only the nearest legitimate sample is considered. Moreover, the average distance from a malicious sample to its nearest legitimate sample is 0 when the value range of malicious samples is covered by legitimate samples. This happens frequently for Boolean and categorical features since their values are limited. Therefore, these kinds of features may not be evaluated precisely.

This study aims to propose an outlier-insensitive adversary-awareness filter feature selection method which can deal with an application with numerical, categorical and Boolean features. The proposed model is called Distribution Based Adversarial Filter Feature Selection (DAFFS). Different from FAFS [19], the distribution of a feature candidate is considered, which reduces the influence of outliers. The generalization ability is quantified by Symmetrical Uncertainty (SU) which measures the relevance between a candidate feature set and class in terms of normalized information gain. On the other hand, Earth Mover's Distance (EMD) [20] defined as the minimum modification from a class to another, represents the cost of attack, *i.e.,* the robustness against adversarial attacks. SU and EMD are able to quantify the characteristic of a feature accurately even though the data type only contains a few values. The proposed feature selection model is then evaluated and compared with the traditional filter method and the existing adversarial filter method [19] in various settings experimentally. The experimental results suggest our proposed method achieves satisfying results, and the data distribution-

based feature evaluation is more stable and flexible.

The rest of the paper is organized as follows. Section II provides a brief introduction on the related concepts of this study, while DAFFS and its generalization ability and security terms are explained in Section III. The experimental results and discussion are given in Section IV. Finally, Section V concludes this study and discusses future works.

II. Related Work

The related concepts, including adversarial learning and feature selection, are briefly introduced in this section. The notation used in this study is also introduced.

A. Adversarial Learning

Due to the popularity of machine learning, its security issues have drawn attention. An adversary who carefully manipulates samples intentionally in order to mislead the the decisions of a system may exist in a security-related application. For example, words of a spam email may be added, removed or obscured in order to evade detection by spam filtering systems. As machine learning methods do not consider the change on data distribution, many studies indicate that the methods do not work well under adversarial attacks. How to build an accurate and robust system in an adversarial environment becomes a critical problem.

A taxonomy of adversarial attacks against machine learning has been defined in [16], [21]–[23]. The attack can be described in three aspects including the influence, violation and specificity. This study focuses on evasion attacks in which an adversary aims to mislead the decisions of a classifier on test samples by crafting their feature values. The manipulation level is limited by the attack strength which is the attack cost. All knowledge, *e.g.,* the model and its parameters, of the targeted classifier is obtained by an adversary, which is the worst scenario to a defender.

Let $\mathbf{x}$ be a test sample and $g(\mathbf{x})$ be a classifier. $\mathbf{x}$ is classified as a malicious class if $g(\mathbf{x}) \geq 0$; otherwise, $\mathbf{x}$ belongs to a legitimate class. Given an attack strength d_{max} defined as the maximum modification on $\mathbf{x}$, an evasion attack is formulated as an optimization problem which aims to minimize $g(\mathbf{x})$ with modification smaller than d_{max}:

$$\mathbf{x}^* = argmin_{\mathbf{x}} g(\mathbf{x}) \quad \text{subject to } d(\mathbf{x}, \mathbf{x}^*) \leq d_{max} \tag{1}$$

where $\mathbf{x}^*$ denotes the attack sample, $d(\mathbf{x}, \mathbf{x}^*)$ represents a distance function between $\mathbf{x}$ and $\mathbf{x}^*$. For example, L_1 or L_2 norms.

Robust learning is one defense against evasion attacks. It aims to increase the cost of evasion by considering robustness in training. Some methods optimize a robustness term explicitly together with the generalization ability when crafting a decision plane [11]. Another kind of methods [7] injects a number of training samples contaminated by an evasion attack into the training set. The trained classifier is focused on classifying attack samples correctly. Increasing the classifier complexity [3], [12] is also an effective way to increase the robustness of a classifier.

Data sanitization is another countermeasure against evasion attacks and works by pre-processing the training set by identifying suspected data. Reject On Negative Impact (RONI) [24] eliminates the samples which have a negative impact to classification accuracy. A classifier is trained with and without a sample. This sample is removed if the accuracy of the classifier without it is higher than the one with it. Data complexity, which measures the difficulty of a classification task, is used as another criterion in sample evaluation [25].

B. Feature Selection

Feature Selection [26]–[28] is the process of identifying a feature subset which contributes the most to the problem. It plays important roles in many applications in order to reduce the complexity and also increase the accuracy. Filter approaches [29]–[31] use the inherent statistical property of a dataset to evaluate the relationship between a feature set and class without relying on a classifier, while feature evaluation criteria used in wrapper approaches [32]–[34] depend on the performance of a classifier. Usually, the time complexity of a filter approach is smaller but the features selected by a wrapper approach achieve better performance. Embedded approaches [35]–[38] avoid the iterative selection by optimizing the classifier and feature set simultaneously.

Some studies [16] have demonstrated that features selected by traditional feature selection may be vulnerable in adversarial environments since only the generalization ability is considered. In order to select discriminative and robust features, several adversarial feature selection methods considering both generalization ability and security (S) have been proposed.

The adversary-aware feature selection method (WAFS) [16], which is a wrapper method, considers an additional security term (S) besides the training error in the objective function. S calculates the expectation of successful attack cost c of samples, defined as $S(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|Y=1)} c(\mathbf{x}^*_{\boldsymbol{\theta}}, \mathbf{x}_{\boldsymbol{\theta}})$, where $\mathbf{x}_{\boldsymbol{\theta}}$ and $\mathbf{x}^*_{\boldsymbol{\theta}}$ are an original sample and its attack sample, $\mathbf{x}$ denotes a set of samples in the distribution of p, Y represents the label, and $Y = 1$ indicates the malicious class. The same as other wrapper methods, one drawback of WAFS is the time complexity.

FAFS [19] quantifies security as the averaged distance from each malicious sample to its nearest legitimate sample. G measures the information gain and S is defined as: $S(\boldsymbol{\theta}) = \sum_{x^+_{\boldsymbol{\theta}} \in M} \min_{x^-_{\boldsymbol{\theta}} \in L} d(x^+_{\boldsymbol{\theta}}, x^-_{\boldsymbol{\theta}})/|M|$, where $x^+_{\boldsymbol{\theta}}$ and $x^-_{\boldsymbol{\theta}}$ represent a malicious and legitimate sample respectively on the selected features of $\boldsymbol{\theta}$ from the malicious (M) and legitimate set (L), and $|M|$ is the cardinality. $\min d(x^+_{\boldsymbol{\theta}}, x^-_{\boldsymbol{\theta}})$ measures the distance between the closest malicious and legitimate samples. As no training is involved in evaluation, its running time is significantly shorter than the wrapper approach. However, this method is sensitive to outliers and may not be suitable for a feature with a few values as the minimum distance is considered.

III. Distribution-Based Adversarial Filter Feature Selection (DAFFS)

As the current adversarial filter feature selection [19] may not evaluate precisely on the features with limited values, *e.g.,* categorical and Boolean features, and also is sensitive to outliers. Distribution-based Adversarial Filter Feature Selection (DAFFS) is proposed and devised in this section. As the distribution of every pair of malicious and legitimate samples are considered in the evaluation on security, the influence of outliers can be reduced. Moreover, the distribution evaluation is suitable to different types of features.

Section III-A discusses the framework of our feature selection method. The measure of generalization ability and security, which are Symmetrical Uncertainty (SU) and Earth Mover's Distance (EMD), are introduced in Sections III-B and III-C. The characteristic of a feature set is discussed in terms of SU and EMD in Section III-D.

A. Proposed Algorithm

Given a dataset $\mathcal{D}$ of a c-class classification containing n number of training samples $(\mathbf{x}^{(i)}, y^{(i)})$ where $i = 1, 2, \ldots, n$. $\mathbf{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, ..., x_m^{(i)}\}$ is a m-dimensional feature vector while $y^{(i)} \in \mathcal{C}$ is the class label, where $\mathcal{C} = \{1, 2, ..., c\}$. A feature subset $\boldsymbol{\theta}$ is selected according to (2), where $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_m\} \in \{0, 1\}^m$ is a m-dimensional feature vector. $\theta_i = 1$ indicates whether the ith feature is chosen; otherwise, $\theta_i = 0$. k-feature selection is formalized as:

$$\theta^* = argmax_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) + \alpha S(\boldsymbol{\theta}) \quad \text{subject to } \Sigma_{i=1}^{m} \theta = k \quad (2)$$

where α is a trade-off parameter, $G(\boldsymbol{\theta})$ and $S(\boldsymbol{\theta})$ measure the generalization ability and security of the candidate features θ evaluated by symmetrical uncertainty and Earth Mover's distance respectively. The detail of both terms will be discussed in detail in Sections III-B and III-C.

To avoid an NP-hard problem, a feature selection can be formulated as Forward Selection Method (FSM) or Backward Elimination Method (BEM) [16], [19] in which a feature is selected and removed iteratively. In this study, FSM is used to demonstrate our proposed model. The selected feature set $\mathbf{T}$ is null and the unselected feature set $\mathbf{U}$ is the full set at the beginning. A feature f with the largest score chosen from $\mathbf{U}$ is added to $\mathbf{T}$ in each iteration until $\mathbf{U} = \emptyset$. The score of f (s_f) is calculated by:

$$s_f = \hat{G}(\mathbf{T} \cup \{f\}) + \alpha \hat{S}(\mathbf{T} \cup \{f\}) \quad (3)$$

where $\hat{G}$ and $\hat{S}$ are the normalized G and S, which are defined as:

$$\hat{M}(\mathbf{T} \cup \{f\}) = \frac{M(\mathbf{T} \cup \{f\}) - \min_M}{\max_M - \min_M} \quad (4)$$

where M represents G or S. $\min_M = \min_{f' \in \mathbf{U}} M(\mathbf{T} \cup \{f'\})$ and $\max_M = \max_{f' \in \mathbf{U}} M(\mathbf{T} \cup \{f'\})$, which represent the minimum and maximum score on all candidates. The detailed algorithm is shown in Algorithm 1.

Algorithm 1 Forward Selection Method of DAFFS

Input: $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^n$: the training dataset with n samples; k: number of features that are needed to select; α: trade-off parameter in feature selection

Output: $\boldsymbol{\theta} \in \{0, 1\}^m$: selected features (1: selected, 0: not selected)

```
T ← ∅, U ← {1, ..., m};
for R = 1, 2, ..., k do
    for each feature f ∈ U do
        Set F ← T ∪ {f};
        Set θ = 0, and then θ_j = 1 for j ∈ F;
        Estimate G_f(θ) and S_f(θ) on D_θ = {x^i_θ, y^i}_{i=1}^n;
    end for
    Normalize G_f(θ) as Ĝ_f(θ)
    Normalize S_f(θ) as Ŝ_f(θ)
    f* ← arg max_f(Ĝ_f(θ) + αŜ_f(θ));
    T ← T ∪ {f*}, U ← U\{f*};
    Set θ^R = 0 and then θ^R_t = 1 for t ∈ T;
end for
Set θ = θ^k;
return θ
```

B. Generalization Ability by Symmetrical uncertainty

Symmetrical Uncertainty (SU) is used to quantify the discrimination ability of a feature subset. It has been found that information gain (IG) biases to features with more varying values [29]. SU is defined as IG normalized by the summation of entropy of the feature and classes in order to eliminate the bias. It is illustrated that SU is a better measure on discrimination ability than IG [39].

Let Y be the class label of the samples. SU for a feature set F is defined as:

$$SU(Y, F) = 2\left[\frac{IG(Y|F)}{H(Y) + H(F)}\right] \quad (5)$$

where $IG(Y|F)$ is the information gain defined as $IG(Y|F) = H(Y) - H(Y|F)$, and H is the entropy. IG measures the uncertainty of Y decreased by using F. IG is normalized by $H(Y) + H(F)$ to reduce the bias. The range of SU is [0,1]. Larger SU indicates F provides more information in classifying Y, *i.e.,* F with larger SU is more discriminative.

C. Security Evaluation by Earth Mover's distance

Earth Mover's distance (EMD) [20], also known as Wasserstein distance, calculates the minimum modification required by changing a distribution to another one. If the distributions of classes are too similar, small manipulation on samples can mislead the decision of a classifier easily. As a result, EMD has been used to quantify the security of a classifier in an evasion attack. A feature space with a larger distance between distributions of classes is preferable.

Given a feature space X, EMD for the distance measurement between the probability density functions of Class A and B (p_A and p_B) is defined as:

$$W(p_A, p_B) = \inf_{\rho \in (p_A, p_B)} \iint_{\mathbf{x}_A, \mathbf{x}_B \in X} \rho(\mathbf{x}_A, \mathbf{x}_B) d(\mathbf{x}_A, \mathbf{x}_B) \mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \tag{6}$$

where $\mathbf{x}_A$ and $\mathbf{x}_B$ are the variables representing the samples in Class A and B, and d denotes a distance function. ρ is a joint density function of p_A and p_B, and satisfies the conditions of $\int_{\mathbf{x}_B \in X} \rho(\mathbf{x}_A, \mathbf{x}_B) d\mathbf{x}_B = p_A$, $\int_{\mathbf{x}_A \in X} \rho(\mathbf{x}_A, \mathbf{x}_B) d\mathbf{x}_A = p_B$, and $\rho(\mathbf{x}_A, \mathbf{x}_B) \geq 0$. EMD can be formulated as as an optimization problem and solved by using the linear programming [40], [41] without any assumption of the sample distribution. Therefore, EMD is able to deal with continuous, discrete and categorical features. The ranges of EMD is from 0 to $+\infty$. $W = 0$ means the distributions of two classes are the same. Larger EMD value means a greater difference between the distributions in feature space.

D. Discussion on Generalization Ability and Security

The measurements of the generalization ability and security, *i.e.,* SU and EMD, used in the proposed method are discussed and illustrated by artificial datasets. EMD evaluates the hardness of manipulation for a successful attack. A larger EMD value means that a larger feature value modification is required for detection evasion, *i.e.,* the system is more secure. Meanwhile, SU indicates that the discriminant ability. Hence, larger SU and EMD values are more preferable. Figure 1 illustrates four datasets with different degrees of generalization ability and security. Red and blue indicate two classes.

Figure 1(a) indicates that two classes of Dataset A are highly separable, and the samples of both classes are far away from each other, *i.e.,* the generalization ability and security are good. Therefore, the values SU and EMD are larger than other datasets. Although the classes are overlapped in Dataset B, samples of two classes are not close to each other. Therefore, its SU is smaller than Dataset A but EMD is still relatively high. In Dataset C, two classes form concentric circles, which can be clearly separated and SU is high. However, the samples between two classes are close to another, *i.e.,* EMD of Dataset C is much smaller than A and B. On the other hand, in Dataset D, the distributions of two classes are nearly overlapped and little modification can be done to move a sample to another class. Therefore, its SU and EMD are also small . The artificial datasets illustrate that SU and EMD are able to measure the generalization ability and security reasonably according to the data distribution.

IV. Experiments

In this section, a number of experiments are carried out to evaluate and compare the features selected by our proposed method and the state-of-the-art methods in the security-related applications in terms of generalization and robustness to evasion attack.

A. Experimental Settings

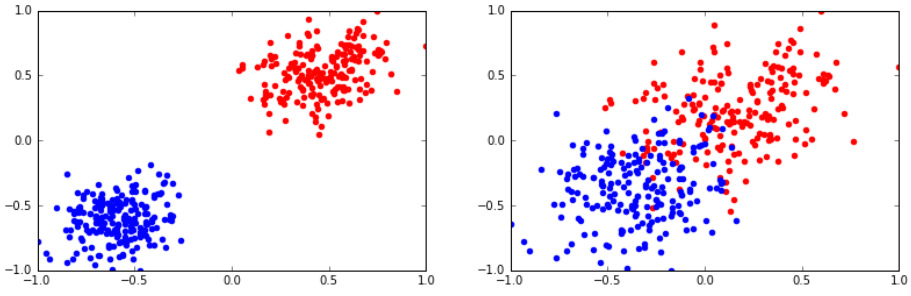
Real security-related applications, including PDF malware (PDF) [42] and TREC 2007 email corpus (SPAM) [43], are adopted in our experiments. PDF dataset has 5993 malicious and 5951 legitimate samples with 114 features, which are in the range of [0,1]. In consideration of the mechanism of PDF, keywords of PDF can only be appended but not be removed. On the other hand, SPAM dataset is first pre-processed as [16], and contains 2500 malicious and 2500 legitimate samples with 500 boolean features representing whether the keywords exist in the samples.

Each experiments are carried out ten times independently. In each experiment, the dataset is randomly separated into half and half for training and testing sets. Preliminary experiments are carried out to determine the parameters for classifiers and feature selection methods aiming at maximizing the accuracy. An SVM with the Gaussian kernel with $C = 0.5$ and $\gamma = 0.1$ is used for PDF, while an SVM with the linear kernel with $C = 1$ is chosen for SPAM. The true positive rate when the false positive rate = 1% (TP@FP=1%) is used to quantify the performance of a classifier with a feature subset since misclassification on legitimate samples (*i.e.,* the false positive rate) as malicious is more harmful in most security-related applications. Our proposed method *DAFFS* is compared with the traditional filter feature selection method with SU (*SU*), *FAFS* [19], and *WAFS* [16]. The evasion attack using the gradient descent algorithm [16] is applied to craft a attack sample. The gradient step size, the maximum number of iterations and the stop condition are set as $t = 0.002$, $m = 1000$ and $\epsilon = 0.01$ respectively. The attack strength defined as the maximum difference between a sample before and after manipulation is set as $[0, 0.05, 0.1, ..., 0.3]$ and $[0, 4, 8, ..., 20]$ in PDF and SPAM respectively. We assume an adversary has full knowledge of the target classifier and its parameters. Moreover, only malicious samples are attacked.

B. Malware Detection in PDF Files

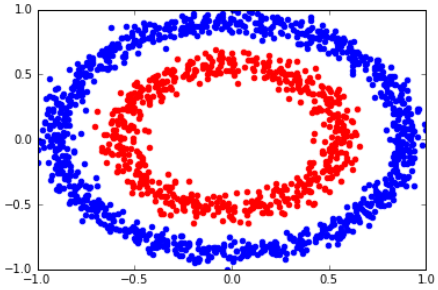
The PDF dataset with real values between 0 and 1 are discussed in this section. TP@FP=1% of *SU*, *FAFS*, *WAFS*, and *DAFFS* with 10, 40, 70, and 100 selected features under attack strengths are shown in Figure 2. Each line represents the average results achieved by one feature selection method, and the range on each point denotes the variance of TP@FP=1% of 10 independent runs.

SU has smaller TP@FP=1% than the other three adversarial feature selection methods in all cases, and the values drop faster with the increase of attack strength, which means the features selected by the adversarial feature selection methods are robust against evasion attacks. When more features are used, the TP rate of *SU* has a great improvement compared with the other three methods under the same attack strength, but less difference between adversarial feature selection methods since many selected features are the same. However, *SU* still performs significantly badly, which means taking a security term into consideration plays an important role in an adversarial environment. Moreover, the four feature selection

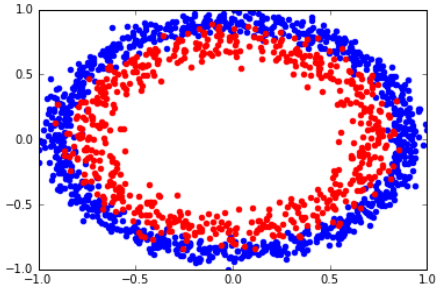


(a) Dataset A (SU =0.2581, EMD =2.1846)

(b) Dataset B (SU =0.1639, EMD =1.1052)



(c) Dataset C (SU =0.2076, EMD =0.4365)



(d) Dataset D (SU =0.1649, EMD =0.2395)

Fig. 1: Four Artificial Datasets with different generalization ability and security.

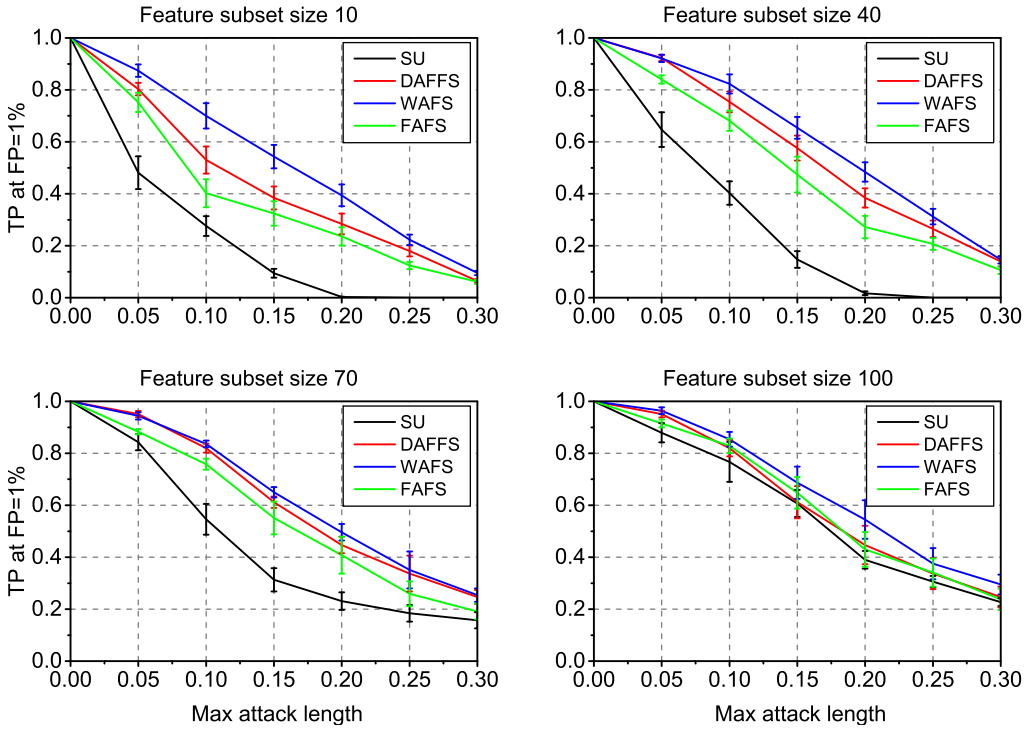


Fig. 2: TP@FP=1% of SVMs on the PDF dataset with 1, 40, 70, and 100 selected features under different attack strengths

methods have all satisfied TP rate without attack, that is, adversarial feature selection method could perform well under and in the absence of an attack.

Our proposed *DAFFS* achieves worse performance than *WAFS* because the features selected by *WAFS* are chosen specifically to the classifier experimentally but *DAFFS* selects the features generally. On the other hand, *DAFFS* outperforms *FAFS* in most cases when the sizes of the selected feature set are 10, 40, and 70. The results indicates that the distribution-based measure used by our study quantifies the security of

features better than the one used in *FAFS*.

C. Spam Detection

The performance of the feature selection methods on Boolean features are evaluated by using the SPAM dataset. Figure 3 shows results of selecting 100, 200, 300 and 400 features in terms of TP@FP=1% under different maximum modified words.

Similar results to the previous experiments on continuous-valued features can be observed. *WAFS* has the best performance among all methods, while TP@FP=1% of *SU* is the lowest. *SU* drops sharply when test samples are attacked, for example, TP@FP=1% decreases from 88% to 2% when at most four out of 100 words can be manipulated. Different from the results in the PDF dataset, the difference of *DAFFS* and *FAFS* decreases significantly with the increase of the size of the selected feature. The results indicate that *FAFS* is not suitable to deal with boolean features when only a few features are considered. Moreover, feature selection improves the robustness of the classifiers in the adversarial environment. In most cases, the TP rate of any method with 300 selected features are higher than the one with 400 selected features under the same attack.

D. Stability to Noisy Environment

The stability of two adversarial filter feature selection methods, *i.e., DAFFS* and *FAFS*, against noise is investigated. The feature and label noise are considered separately in our settings. The feature noise is generated by deviating feature values of 3% samples by the Gaussian noise with the variance 0.2, while the labels of 3% samples are flipped for the label noise in this experiment.

Stability index [44], which evaluates the consistency of feature subsets selected from two datasets, is adopted as measurement. The range of the stability index value is $[-1, 1]$. A larger value means the feature selection method is more stable, *i.e.,* the selected feature subsets are similar when the dataset is noisy, vice versa. For a dataset with m features and k features will be selected, assume the number of intersecting features between feature subset A and B is r, the stability index of A and B is defined as:

$$I_c(A, B) = (rm - k^2)/[k(m - k)], \tag{7}$$

where A and B denote the feature subsets selected from the original and contaminated datasets. The results are shown in Figure 4.

As mentioned above, *SU* is not suitable for feature selection in an adversarial environment, while *WAFS* is time consuming. Therefore, the stability of the only two filter feature selection methods, including *DAFFS* and *FAFS*, are analyzed in a noisy environment. The index values of *DAFFS* are significantly higher than *FAFS* in most cases, which indicates *DAFFS* is much more stable than *FAFS*. It confirms that considering the closest malicious sample in *FAFS* is more noise sensitive than the distribution-based measure in our method. In addition, *FAFS* has worse performance on selecting Boolean than real-valued features in terms of stability. The stability index values of *FAFS* are much smaller in the Spam dataset than the PDF dataset. However, our *DAFFS* has similar performance in both datasets.

V. Conclusion

Feature selection is an essential pre-processing step in machine learning. Although the existing adversarial filter feature selection method is able to select discriminative and robust features in an adversarial environment, it is noise sensitive and not suitable for Boolean features. This study applies distribution-based measures, *i.e.,* Earth Mover's Distance and Symmetric Uncertainty, to quantify the security and generalization ability of a feature set. The experimental results confirm that the feature subsets selected by our proposed method are more discriminative, robust to adversarial attacks, feature and label noise. Moreover, the distribution-based measures are suitable to evaluate Boolean and real-valued features. One possible future work is to select effective features on a contaminated training set. Some samples may be manipulated in an adversarial environment and will affect the selected feature. How to reduce the influence by the contaminated samples in selecting features is a key research problem.

Acknowledgment

This paper is supported by the Natural Science Foundation of Guangdong Province, China (No. 2018A030313203) and the Fundamental Research Funds for the Central Universities (No. 2018ZD32).

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Th European Conference on Machine Learning & Knowledge Discovery in Databases*, 2013.

[2] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli, "Support vector machines under adversarial label contamination," *Neurocomputing*, vol. 160, pp. 53–62, 2015.

[3] Z. Khorshidpour, S. Hashemi, and A. Hamzeh, "Learning a secure classifier against evasion attack," in *IEEE International Conference on Data Mining Workshops*, 2017.

[4] A. Chernikova, A. Oprea, C. Nita-Rotaru, and B. Kim, "Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction," *arXiv preprint arXiv:1904.07370*, 2019.

[5] D. Hitaj and L. V. Mancini, "Have you stolen my model? evasion attacks against deep neural network watermarking techniques," *arXiv preprint arXiv:1809.00615*, 2018.

[6] H. Xiao, T. Stibor, and C. Eckert, "Evasion attack of multi-class linear classifiers," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2012, pp. 207–218.

[7] L. Chen, Y. Ye, and T. Bourlai, "Adversarial machine learning in malware detection: Arms race between evasion attack and defense," in *2017 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2017, pp. 99–106.

[8] P. P. Chan, Z. Lin, X. Hu, E. C. Tsang, and D. S. Yeung, "Sensitivity based robust learning for stacked autoencoder against evasion attack," *Neurocomputing*, vol. 267, pp. 572–580, 2017.

[9] Y. Shi and Y. E. Sagduyu, "Evasion and causative attacks with adversarial deep learning," in *MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM)*. IEEE, 2017, pp. 243–248.

[10] G. Piccioli, E. De Micheli, P. Parodi, and M. Campani, "Robust method for road sign detection and recognition," *Image and Vision Computing*, vol. 14, no. 3, pp. 209–223, 1996.

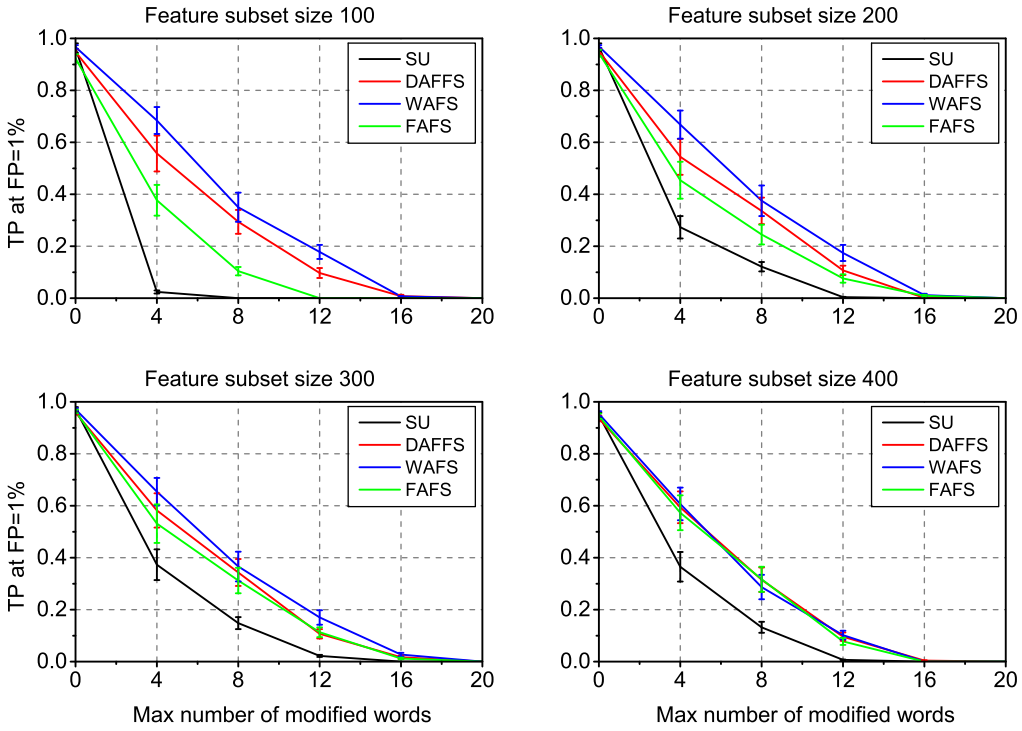
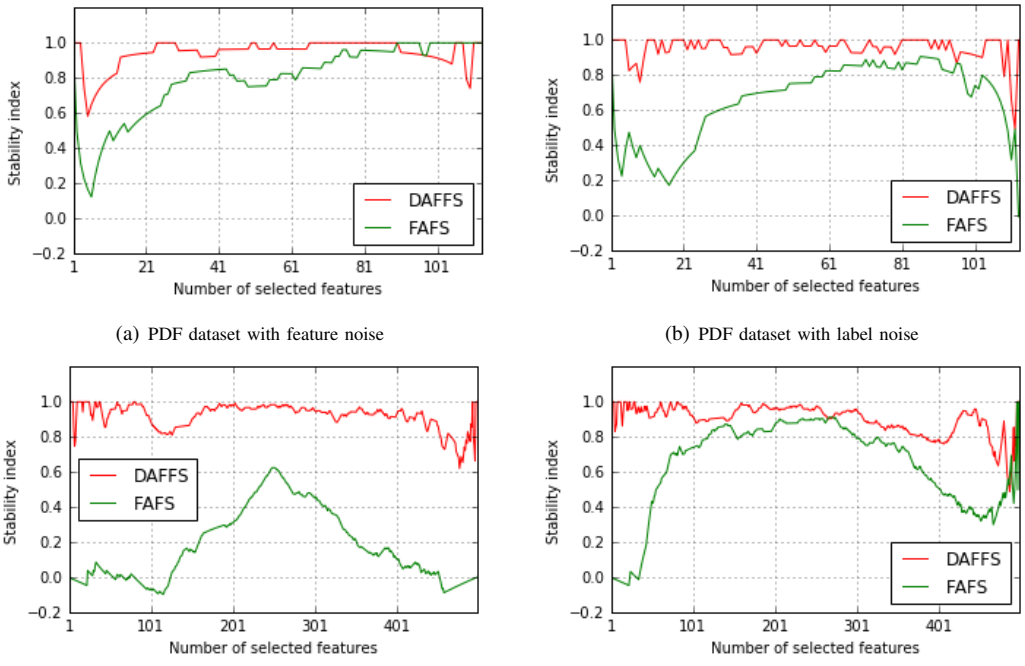


Fig. 3: TP@FP=1% of SVMs on the SPAM dataset with 100, 200, 300, and 400 selected features under different attack strengths



(a) PDF dataset with feature noise

(b) PDF dataset with label noise

(c) Spam dataset with feature noise

(d) Spam dataset with label noise

Fig. 4: Stability index of the feature selection results on the original and contaminated datasets of *FAFS* and *DAFFS*.

[11] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Tenth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, 2004.

[12] B. Biggio, G. Fumera, and F. Roli, "Multiple classifier systems for robust classifier design in adversarial environments," *International Journal of Machine Learning & Cybernetics*, vol. 1, no. 1-4, pp. 27–41, 2010.

[13] Z. He, J. Su, M. Hu, G. Wen, and Z. Fei, "Robust support vector machines against evasion attacks by random generated malicious samples," in *2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, 2017.

[14] H. Kwon, Y. Kim, K.-W. Park, H. Yoon, and D. Choi, "Multi-targeted adversarial example in evasion attack on deep neural network," *IEEE Access*, vol. 6, pp. 46 084–46 096, 2018.

[15] F. Khalid, M. A. Hanif, S. Rehman, J. Qadir, and M. Shafique, "Fademl: Understanding the impact of pre-processing noise filtering on adversarial machine learning," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019.

[16] F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung, and F. Roli, "Adversarial feature selection against evasion attacks," *IEEE transactions on cybernetics*, vol. 46, no. 3, pp. 766–777, 2015.

[17] Z. Katzir and Y. Elovici, "Quantifying the resilience of machine learning classifiers used for cyber security," *Expert Systems with Applications*, vol. 92, p. S0957417417306590, 2017.

[18] G. Ditzler and A. Prater, "Fine tuning lasso in an adversarial environment against gradient attacks," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–7.

[19] M. Wu and Y. Li, "Adversarial mrmr against evasion attacks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–6.

[20] M. Werman, S. Peleg, and A. Rosenfeld, "A distance metric for multidimensional histograms," *Computer Vision Graphics & Image Processing*, vol. 32, no. 3, pp. 328–336, 1985.

[21] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[22] J. D. Tygar, "Adversarial machine learning," in *Acm Workshop on Security & Artificial Intelligence*, 2011, pp. 4–6.

[23] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli, "Is feature selection secure against training data poisoning?" in *International Conference on Machine Learning*, 2015, pp. 1689–1698.

[24] N. B. A, "Behavior of machine learning algorithms in adversarial environments," *California University Berkeley, Department of Electrical Engineering and Computer Science*, 2010.

[25] P. P. K. Chan, Z. M. He, H. Li, and C. C. Hsu, "Data sanitization against adversarial label contamination based on data complexity," *International Journal of Machine Learning & Cybernetics*, vol. 9, no. 6, pp. 1039–1052, 2018.

[26] H. L. M. Dash, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[27] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. 6, pp. 1157–1182, 2003.

[28] P. Hanchuan, L. Fuhui, and D. Chris, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[29] Y. Lei and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Twentieth International Conference on International Conference on Machine Learning*, 2003.

[30] P. A. Estévez, T. Michel, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[31] L. T. Vinh, N. D. Thang, and Y. K. Lee, "An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information," in *IEEE/IPSJ International Symposium on Applications & the Internet*, 2010.

[32] H. Liu and R. Setiono, "Feature selection and classification - a probabilistic wrapper approach," in *International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems*, 1996.

[33] M. M. Kabir, M. M. Islam, and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, vol. 73, no. 16, pp. 3273–3283, 2008.

[34] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Feature subset selection problem using wrapper approach in supervised learning," *International Journal of Computer Applications*, vol. 1, no. 7, pp. 13–17, 2010.

[35] Y. Kim and J. Kim, "Gradient lasso for feature selection," in *International Conference on Machine Learning*, 2004.

[36] B. J. Marafino, B. W. John, and D. R. Adams, "Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to icu risk stratification from nursing notes," *Journal of Biomedical Informatics*, vol. 54, no. C, pp. 114–120, 2015.

[37] S. Paul and P. Drineas, "Feature selection for ridge regression with provable guarantees," *Neural computation*, vol. 28, no. 4, pp. 716–742, 2016.

[38] S. Zhang, D. Cheng, R. Hu, and Z. Deng, "Supervised feature selection algorithm via discriminative ridge regression," *World Wide Web*, no. 1, pp. 1–18, 2017.

[39] K. J. T., "Information gain and a general measure of correlation," *Biometrika*, no. 1, p. 1, 1983.

[40] M. Kojima, S. Mizuno, and A. Yoshise, "A primal-dual interior point algorithm for linear programming," in *Progress in mathematical programming*. Springer, 1989, pp. 29–47.

[41] V. Herrmann, "Wasserstein gan and the kantorovich-rubinstein duality," https://vincentherrmann.github.io/blog/wasserstein/, accessed February 4, 2017.

[42] D. Maiorca, I. Corona, and G. Giacinto, "Looking at the bag is not enough to find the bomb: An evasion of structural methods for malicious pdf files detection," in *Acm Sigsac Symposium on Information*, 2013.

[43] G. V. Cormack and T. R. Lynam, "Trec 2005 spam track overview," in *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, November 15-18, 2005*, 2005.

[44] L. I. Kuncheva, "A stability index for feature selection," in *Conference on Iasted International Multi-conference: Artificial Intelligence & Applications*, 2007.