Defense Against Poisoning Attack via Evaluating Training Samples Using Multiple Spectral Clustering Aggregation Method

Wentao Zhao[1], Pan Li[1, *], Chengzhang Zhu[1, 2], Dan Liu[1] and Xiao Liu[1]

Abstract: The defense techniques for machine learning are critical yet challenging due to the number and type of attacks for widely applied machine learning algorithms are significantly increasing. Among these attacks, the poisoning attack, which disturbs machine learning algorithms by injecting poisoning samples, is an attack with the greatest threat. In this paper, we focus on analyzing the characteristics of positioning samples and propose a novel sample evaluation method to defend against the poisoning attack catering for the characteristics of poisoning samples. To capture the intrinsic data characteristics from heterogeneous aspects, we first evaluate training data by multiple criteria, each of which is reformulated from a spectral clustering. Then, we integrate the multiple evaluation scores generated by the multiple criteria through the proposed multiple spectral clustering aggregation (MSCA) method. Finally, we use the unified score as the indicator of poisoning attack samples. Experimental results on intrusion detection data sets show that MSCA significantly outperforms the K-means outlier detection in terms of data legality evaluation and poisoning attack detection.

Keywords: Poisoning attack, sample evaluation, spectral clustering, ensemble learning.

1 Introduction

In big data era, machine learning is becoming one of the most popular techniques in many applications because of its excellent performance. For example, in image recognition [Lingyun, Xiaobo, Jiaohua et al. (2018)], many machine learning algorithms have achieved higher recognition accuracy compared with human [Makili, Vega, Dormido-Canto et al. (2011)]. Besides, those algorithms also show significant successes in other domains, such as speech recognition [Hinton, Deng, Yu et al. (2012)], intrusion detection system (IDS) [Tsai, Hsu, Lin et al. (2009)], financial prediction [Wei, Liang and Longbing (2015)], web content analysis [Lingyun, Yan, Wei et al. (2018)] and data analytics [Chengzhang, Longbing, Qiang et al. (2018); Lingyun, Guohan, Qian et al. (2018)], etc.

However, recent researches indicate that many attacks can destroy the application of

[1] College of Computer, National University of Defense Technology, Changsha, 410073, China.

[2] Faculty of Engineering and Information Technology, University of Technology Sydney, 2007, Australia.

[*] Corresponding Author: Pan Li. Email: lipan16@nudt.edu.cn.

CMC. doi:10.32604/cmc.2019.05957

machine learning [Liu, Li, Zhao et al. (2018)]. Basically, these attacks can be classified into two categories: exploratory attack and causative attacks [Barreno, Nelson, Sears et al. (2006)]. Exploratory attack mainly destroys the learning model performance during the prediction stage. For example, the adversarial sample crafting methods proposed by [Nguyen, Yosinski and Clune (2015); Carlini, Mishra, Vaidya et al. (2016)] can easily manipulate a well-trained deep neural network (DNN). Causative attack mainly occurs in the model training stage. A typical causative attack method is the poisoning attack. Poisoning attack manipulates the learning model by modifying the features or labels of training data or injecting poisoning data that is similar yet has different distribution to training data. As evidenced by [Biggio, Fumera, Roli et al. (2012); Pan, Qiang, Wentao et al. (2018); Pan, Wentao, Qiang et al. (2018)], poisoning attack seriously destroys the performance of learning models that invalidates various applications in multiple scenarios.

Recently, many efforts have been paid on the defense technologies for poisoning attack. Existing methods reduce the impact of poisoning attack by adopting data sanitization and introducing robust learning algorithms. Data sanitization improves the quality of training data via filtering suspicious poisoning data [Nelson, Barreno, Chi et al. (2008); Laishram and Phoha (2016); Paudice, Munozgonzalez and Lupu (2018)]. In contrast, robust learning algorithms tolerant poisoning data through well-designated models. Although the above methods have achieved remarkable performance, they may fail to tackle the following challenges. First, most defense techniques may have a high false detection rate. These methods are sensitive to their hyper-parameters. Without well-tuned hyper-parameters, they are likely to classify the legal samples as poisoning data, and thus, cause the high false detection rate. Second, most defense techniques only suit for a certain kind of attacks. Their scalability and generalization performance should be further improved.

In this paper, we propose a novel poisoning attack defense technique. The proposed technique detects poisoning samples via a multiple spectral clustering aggregation (MSCA) method, which evaluates training samples from multiple views and provides a more robust solution. The main contributions of this paper are summarized as follows:

- We propose an evaluation method to quantify the legitimacy of training samples. This method combines spectral clustering with similarity metric to provide reliable results.
- We further integrate multiple spectral clustering results per different types of similarity metric and various number of clustering centers to form a robust sample evaluation method.
- The MCSA evaluation results can be used to implement a variety of poisoning attack defense technologies under different assumptions catering for specific data characteristics.

The rest of this paper is organized as follows. Section 2 introduces the related knowledge about the adversarial model and the adversarial sample evaluation. Section 3 details the spectral clustering-based evaluation method. Section 4 introduces the proposed multiple spectral clustering aggregation method. Section 5 shows the simulation results by using the MSCA method to remove the suspicious poisoning samples. Section 6 concludes the paper.

2 Preliminaries

2.1 Adversary model

Before studying a type of attack, we should make an assumption about the attacker. Wittel et al. [Wittel and Wu (2004)] first considered the attacker's knowledge when studying the evasive attack of the spam filtering system. Later, some researchers suggested that it is necessary to consider the adversary goal and the adversary capability [Barreno, Nelson, Searset et al. (2006)]. These three aspects further formed the concept of the adversary model, and its specific meaning is as follow:

- **Adversary goal.** Adversary goal is the final effect that the adversary wants to achieve. In this paper, the goal of the attacker is crafting poisoning data to destroy learning models.
- **Adversary knowledge.** Before launching attack, adversary needs some information related to the targeted learning models. In this paper, we suppose the adversary know the whole training data, or the algorithms, even the concrete parameters of targeted models. In this assumption, the attacker can craft various poisoning samples, which can further verify the performance of proposed defense method.
- **Adversary capability.** Besides the adversary knowledge, the adversary should have ability to launch attack. For the poisoning attack, previous works mainly contain two types. One assumes the adversary can change the labels or the features of training data [Zhao, An, Gao et al. (2017); Biggio, Nelson and Laskov (2012)]. The other assumption is that the adversary can inject adversarial samples into training data when the learning models are retraining [Rubinstein, Nelson, Huang et al. (2009); Kloft and Laskov. (2010)]. In this paper, we study the defense technology against the poisoning samples crafted by the latter assumption.

2.2 Sample legitimacy evaluation

2.2.1 Poisoning samples

According to various adversary capabilities, the poisoning samples are mainly crafted by two ways. The first kind of poisoning sample is generated by modifying the features or the labels of existing training data. The second kind of poisoning sample is from the new injecting samples during the model retraining process. Both of the two kinds of poisoning samples have the same attack goal of drifting the initial distribution of training data. Referring to [Pan, Qiang, Wentao et al. (2018)], the poisoning samples have two features.

- The sample should be recognized as legal sample by the classifier. This means that the poisoning samples are not easily identified as abnormal points.
- The sample can change the original training data distribution, which can further affect the performance of the model trained from the training data.

In general, the poisoning samples can cause data drifting of the training data without arousing the suspicion of the classifier. Besides, it can also affect the performance of the targeted learning model.

2.2.2 Sample legitimacy evaluation

Using the poisoned training data without any distinction can seriously decrease the performance of learning model. So, it is critical to properly evaluate the training data. From above introduction, the poisoning data can be defined as the data which can destroy the distribution of training data. In the contrary, legal data represents the initial training data without any contamination by the poisoning data.

Therefore, the sample legitimacy evaluation can quantify the significance or confidence of the training data. So, we define a metric, named *Legitimacy Coefficient* (*LC*), to measure the legality of training samples. Higher *LC* represents that the sample is more closed to the true sample distribution, while lower *LC* shows that the sample is farer away from the true sample distribution.

3 Single spectral clustering for sample evaluation

Combining with the above description of the poisoning samples' characteristics, we first design a single spectral clustering evaluation method to evaluate the *LC* value of training samples. Furthermore, we use the ensemble strategy to integrate multiple spectral clustering learners and then propose a multiple spectral clustering aggregation method, which can be used to evaluate the legitimacy of training data. In this section, we detail the single spectral clustering evaluation method.

3.1 Spectral clustering

Spectral clustering is a clustering algorithm developed from graph theory. The main idea is treating all samples as points in high-dimensional space and connecting them with edges to form an undirected weight map. The weight w_{ij} of each edge in the graph can be used to represent the distance relationship between sample point x_i and sample point x_j. The edge weight of the two sample points will be high if the distance of the two samples is far, and vice versa. The clustering can be realized by cutting the undirected graph, which can ensure the sum of edge weights between two independent sub graphs is as low as possible, and the sum of edge weights inside the sub graph is as high as possible.

The initial spectral clustering can achieve good clustering performance, and the clustering result can provide a benchmark reference for evaluating the training sample. According to [Yu and Shi (2003)], the spectral clustering result is decided by three aspects:

- The computation way of adjacency metric W. In spectral clustering, the edge weight w_{ij} can be used to measure the relationship between two samples. Considering the edge weights of all training samples, we can get the adjacency metric W. As for a data set with n samples, W can be represented by Eq. (1).

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}, \tag{1}$$

where w_{ij} refers to the edge weight between i th sample and j th sample.

- The standard graph laplace metric $L = W - D$, where W is adjacency metric and D refers to degree metric, which can be formulated as Eq. (2).

$$D = \begin{bmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{bmatrix}, \tag{2}$$

where $d_i = \sum_{j=1}^{n} w_{ij}$ represents the sum of edge weight w_{ij} between i th sample and the other samples in the whole data set.

- The clustering way. The last step in spectral clustering is using clustering method such as K-means to deal with the standard feature metric F .

3.2 Single spectral clustering sample rating method

Based on spectral clustering method, we further utilize the clustering result and similarity metric W to realize the evaluation of training data. The specific process of evaluating on each sample is shown in Algorithm 1, v_{xi} represents the *LC* value of sample x_i .

Algorithm 1：Single spectral clustering for sample evaluation

```
Input： Sample set $X = \{x_1, x_2, ..., x_n\}$ with label $\mathbf{p}$ , The generating way of similarity metric $\mathbf{G}$, Class number $\mathbf{k}$
Output：Rating scores $V = \{v_{x1}, v_{x2}, ..., v_{xn}\}$
Initialize: $X_{syn} = \varnothing$
According to $\mathbf{G}$, generate similarity metric $W$
Divide $X$ into $\mathbf{k}$ groups of samples $\{D_1, D_2, ...D_k\}$ using spectral clustering
Choose $D_i$ as the legal sample set $D_{normal} = \{x_i, ..., x_m\}$, other class $D_j (j \neq i)$ as the candidate sample sets, $D_i$ is the closest class to the centre of all samples
for $i = 1$ to $n$ do
    Compute $S_i = \sum_{j=1}^{m} w_{ij}$ , where $w_{ij}$ is the similarity between $x_i$ and $x_j$
end for
Sort $S = \{S_{x1}, ..., S_{xn}\}$ in descending order, the index of $x_i$ is the score for $x_i$
```

Higher value represents higher confidence of sample's legitimacy. In the contrary, lower score refers to lower confidence of sample's legitimacy. The score for each sample is determined by two factors. First, the selection of similarity measure for spectral clustering can be a gaussian kernel function, polynomial kernel function or sigmoid kernel function. Different kernel functions can fit different data distributions. Second, the selection of legal samples class decides the benchmark of basic legitimacy evaluating, which further affect the *LC* value of samples. Besides, the way generating adjacency matrix and the basic clustering way determine the clustering results, which further affect the legal class selection.

4 Multiple spectral clustering aggregation for sample evaluation

For a data set, kernel function type (gaussian kernel, polynomial kernel, sigmoid kernel) and the number of cluster results can influence the evaluation of sample legitimacy. In order to reduce the parameters sensitivity of the algorithm, we combine spectral clustering with ensemble learning to achieve more robust sample legitimacy evaluation.

4.1 Ensemble strategy

The main idea of ensemble learning is integrating multiple base learners into a strong learner through some combination strategies. The performance of the integrated strong learner is more robust than that of each base learner. In ensemble learning, the combination strategy of the base learners directly affects the performance of the integrated strong learner. We simply describe three common combination strategies [Dietterich (2000)] as follows:

- **Averaging strategy:** This is the most common combination strategy for basic learners, including the simple average method and the weighted average method.
- **Voting strategy:** Assuming that the output of each sample is predicted from n learners, the voting method judges the output of the sample by counting the labels predicted by these learners. There are three main ways: the majority of voting methods, the relative majority voting method and the weighted voting method.
- **Learning strategy:** The averaging strategy and voting strategy may not be suitable in some cases. Therefore, it is necessary to construct a nonlinear relationship through learning method to integrate the base learners.

4.2 Average ensemble strategy for multiple spectral clustering aggregation

In this section, we used the averaging strategy as an example to design an aggregation method. The specific process is shown in algorithm 2, MSCA uses the sum of all scores rated by various spectral clustering learners as the *LC* value of the sample evaluated by MSCA. After that, the defender can properly combine the *LC* value with specific defense method to defense against poisoning attack. For example, the defender can adopt the idea of data sanitization. A simple method is removing the samples with the lower *LC* values, which are classified as suspicious poisoning data by MSCA. Besides, based on the value of sample legitimacy provided by MSCA, the defender can further adopt bagging or boosting strategy to ensemble various learners. The *LC* value can be used to adjust the weights of the samples in the training process, which can reduce the impact of poisoning data and achieve more robust machine learning algorithm.

Algorithm 2：Multiple spectral clustering aggregation (MSCA)

```
Input：Sample set $X = \{x_1, x_2, ..., x_N\}$ with label $\mathbf{p}$, the set of generating way of similarity metric $\{G_1, G_2, ..., G_m\}$, the set of clustering number $\{k_1, k_2, ..., k_n\}$
Output：Rating scores $V = \{v_{x1}, v_{x2}, ..., v_{xN}\}$
1. Initialize: $X_{syn} = \varnothing$
2. for $i = 1$ to $n$ do
3.     Choose the number of clustering $k_j$
4.     Input $G_i, k_j, X$ to algorithm 1, get the score $V^{ij}$
5. end for
6. Compute the sum of scores $V_{xp} = \sum_{i_1}^{m} \sum_{j=1}^{n} v_{xp}^{ij}$ for sample $x_p (p = 1, ..., N)$
7. $V = \{v_{x1}, ..., v_{xN}\}$
```

5 Experiment

In this section, we evaluate the performance of the proposed defense method by extensive experiments described as follows: Firstly, we introduce the experimental setting in this paper, including the targeted poisoning samples crating methods and the specific parameters setting of proposed MSCA method. Then we visually show the performances of various methods defense against BEBP poisoning method on synthetic data set. Finally, we compare the proposed method with other defense methods on two real data sets to further demonstrate its stability and effectiveness in the assessment of training data.

5.1 Experimental setting

5.1.1 Poisoning methods

For the poisoning sample crafting methods, we chose two typical poisoning methods, i.e., Batch-EPD Boundary Pattern (BEBP) [Pan, Qiang, Wentao et al. (2018)] and Centre-drifting Boundary Pattern (CBP) [Pan, Wentao, Qiang et al. (2018)]. They can effectively craft boundary pattern data, which can serve as poisoning data to seriously destroy the performance of six different machine learning models.

BEBP This method first randomly divides the training data set into several groups, and then calculates the edge pattern data and corresponding normal vectors of each group data using edge pattern detection algorithm. After that, by pushing the edge pattern data towards their normal vectors, we can get pushed data, which can be further selected by boundary pattern detection algorithm to get the boundary pattern data.

CBP Similar to BEBP, CBP also use the boundary pattern detection method to select the pushed data. The difference is that CBP pushes the whole training data toward the same center vector to get pushing data. And the center vector can be easily obtained by calculating the vector between the two class centres.

Besides, the method in Pan et al. [Pan, Wentao, Qiang et al. (2018)] can launch poisoning attack with weak adversary model, while the attacker should have a strong adversary model in Pan et al. [Pan, Qiang, Wentao et al. (2018)]. For this paper mainly study the defense technology, we suppose that targeted models are in worse case, which means that the attacker has enough knowledge about targeted learning models.

5.1.2 Parameters setting

For the parameters setting of BEBP and CBP, we kept the setting in [Pan, Qiang, Wentao et al. (2018); Pan, Wentao, Qiang et al. (2018)], wherein the poisoning ratio and the poisoning round are set as 0.07 and 5, respectively.

As for the proposed MSCA, we selected six base spectral clustering learners with different kernel types and clustering centers. Specifically, we set the numbers of clustering centre as three and six, and each clustering centre number is combined with three kernel types as follows:

- Gaussian kernel:

$$k(x_i,x_j)=exp(-\|x_i-x_j\|^2/\sigma)$$

- Sigmoid kernel:

$$k(x_i,x_j)=\tanh(\alpha x_i y_i+c)$$

- Polynomial kernel:

$$k(x_i,x_j)=(\alpha x_i y_i+c)^d$$

To focus on evaluation method itself, we simply used the default kernel parameters in the sklearn tool.

5.2 Experiment on synthetic data set

In order to visualize the defense performance of proposed method, we compared the proposed MSCA method with K-means. Meanwhile, the assess result of each base spectral clustering learner is also performed on the synthetic data set.

As shown in the Fig. 1, we can see that the detection results of suspicious poisoning data using various defense methods regarding the poisoning points caused by BEBP method. Obviously, the MSCA method performs better than K-means method and its performance is more stable and effective than single spectral clustering method on the assessment of samples' legitimate.

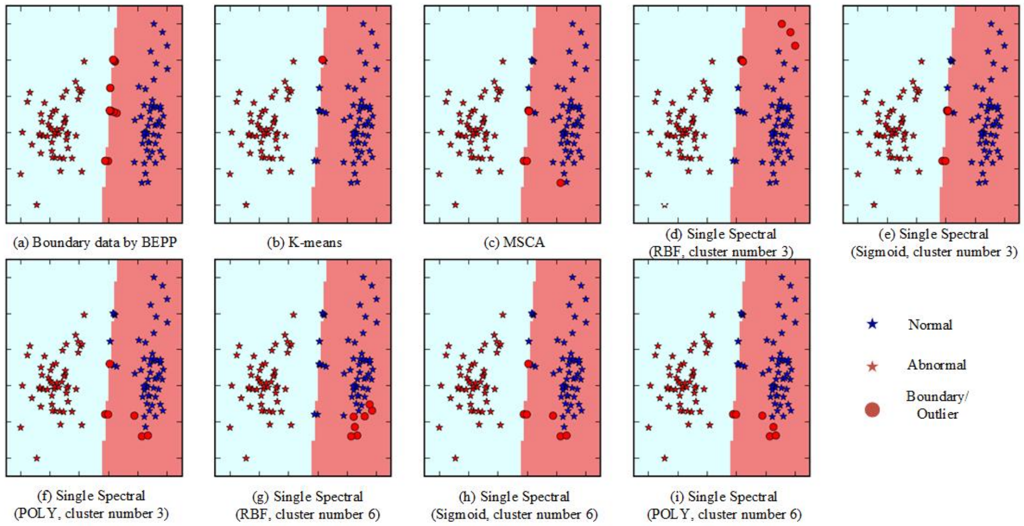
(a) Boundary data by BEPP
(b) K-means
(c) MSCA
(d) Single Spectral (RBF, cluster number 3)
(e) Single Spectral (Sigmoid, cluster number 3)
(f) Single Spectral (POLY, cluster number 3)
(g) Single Spectral (RBF, cluster number 6)
(h) Single Spectral (Sigmoid, cluster number 6)
(i) Single Spectral (POLY, cluster number 6)

Figure 1: The comparison of various defense methods on synthetic data set (The Fig. (a) represents the poisoning effect caused by BEBP, the red round points refer to the boundary poisoning data. The Figs. (b)-(i) represent the defense performance of various defense methods, where the red round points refer to the illegal points detected by various defense methods.)

5.3 Experiments on intrusion detection data sets

Referring to previous work [Pan, Wentao, Qiang et al. (2018)], poisoning attack via injecting poisoning data always occurs in the retraining process of learning model. Network intrusion detection system is a typical domain whose model needs be retrained. So in this section, we evaluate the performance of proposed method by simulating defense experiment on two network intrusion detection data sets.

5.3.1 Performance metric

Regarding an IDS system, accuracy is the primary performance metric. Hence, we adopt accuracy in this paper to evaluate the performance reduction of machine learning-based IDSs under the previous poisoning attack methods. The accuracy (*ACC*) is defined by Eq. (3):

$$ACC = \frac{TP + TN}{TP + TN + FN + FP}, \tag{3}$$

where true positive (TP) is the number of truly abnormal samples that are classified as abnormal ones by IDSs, true negative (TN) means the number of truly normal samples that are treated as normal ones, false positive (FP) refers to the number of truly normal samples classified as abnormal ones, and false negative (FN) represents the number of truly abnormal samples classified as normal ones.

Table 1: Sample distribution of the randomly selected data regarding the NSL-KDD

	Normal	Probing	DOS	U2R	R2L
Training data	2000	300	3790	32	350
Evaluating data	2000	500	3900	20	400

5.3.2 Data sets

To demonstrate the performance of the proposed defense method, we chose two public intrusion detection data sets.

NSL-KDD This data set is a revised version of KDDCUP99, which is a well-known benchmark data set for evaluating the performance of IDSs. NSL-KDD contains five categories of samples (one normal and four abnormal). Moreover, each sample has 41 features.

Kyoto 2006+ This data set proposed in Song et al. [Song, Takakura, Okabe et al. (2011)] is another famous intrusion detection data set. This data set has been collected from honeypots and regular servers that are deployed at the Kyoto University since 2006. Moreover, Kyoto 2006+ contains three types of samples, i.e., normal, known attack and unknown one, and each sample has 24 features.

Referring to Pan et al. [Pan, Qiang, Wentao et al. (2018); Pan, Wentao, Qiang et al. (2018)], we randomly selected training samples and evaluation samples from NSL-KDD and Kyoto 2006+. The sample distribution of training data and evaluating data selected from NSL-KDD shows in Tab. 1. Similarly, we randomly selected 13292 samples from the traffic data collected during 27-31, August 2009 regarding the Kyoto 2006+ data set, including 6472 samples as training data and 6820 samples as evaluating data.

Targeted Models From the previous work we can know that the SVM algorithms were fragile facing the boundary pattern poisoning samples [Pan, Qiang, Wentao et al. (2018)].

Therefore, we focus on the defense methods towards SVM algorithms. Specifically, we select three typical SVM algorithms, including SVM with a radial basis function kernel (SVM-RBF), SVM with a linear kernel (SVM-linear) and SVM with a sigmoid kernel (SVM-Sigmoid).

5.3.3 Experimental results

According to above experimental setting, we evaluated the *ACC* of various defense methods, including *K*-means clustering with outlier removal, the proposed MSCA method and its base spectral clustering with outlier removal, respectively. Meanwhile, the targeted poisoning samples are generated by BEBP and CBP, and both of the two kinds of poisoning methods are performed on NSL-KDD and Kyoto 2006+.

Fig. 2 and Fig. 3 describe the *ACC* changing with no defense, K-means and MSCA method under 5 round poisoning attack using BEBP and CBP on NSL-KDD data set. The No defense method represents the model performance without any defense measure, while *K*-means refers to the *ACC* of model using the *K*-means method to remove the outlier data.

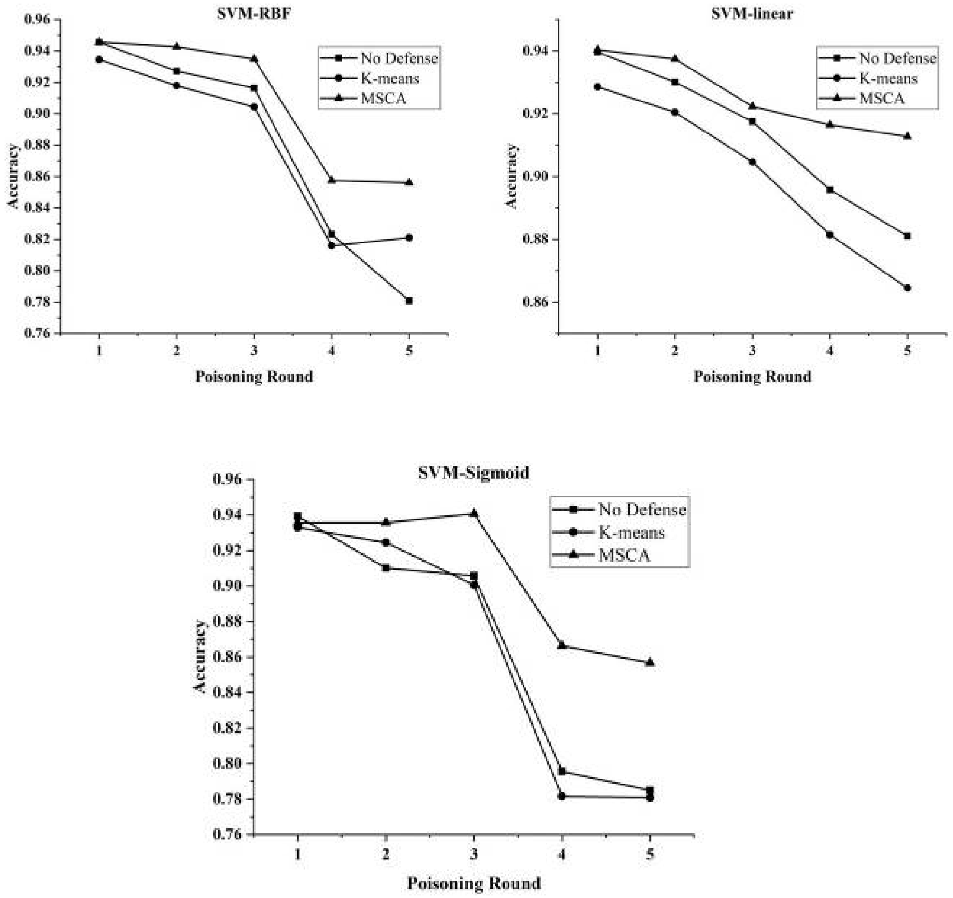

Figure 2: The comparison of *ACC* with various methods defensing against poisoning data crafted by BEBP on NSL-KDD data set

Tab. 2 and Tab. 3 show the results of different methods defensing against 5 round poisoning attack using BEBP and CBP on Kyoto 2006+ data set. Specially, we used Means, Best and Worst in Tab. 2 and Tab. 3 to represent the average *ACC*, the best *ACC* and the worst *ACC* among the six base spectral clustering with outlier removal, respectively.

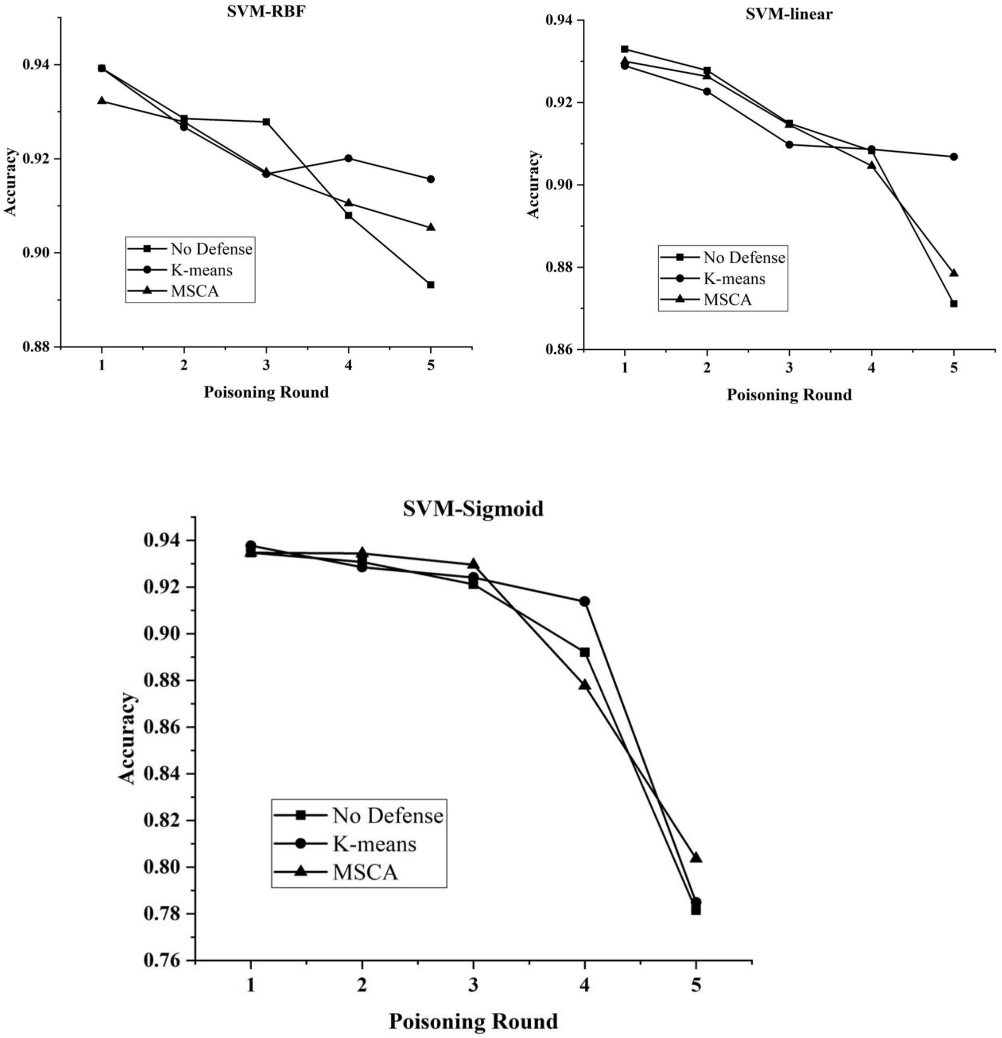

Figure 3: The comparison of ACC with various methods defensing against poisoning data crafted by CBP on NSL-KDD data set

From the above experimental results, we can see that proposed MSCA method can effectively defense the poisoning attack. Specifically, we can compare the defense performance from three aspects.

- As for the performance of defense against different poisoning samples, MSCA can improve the performance of targeted models under poisoning attack of BEBP and CBP both on NSL-KDD and Kyoto 2006+. It is noteworthy that MSCA performs better with the increasing of poisoning rounds. This means that the proposed method can keep the stable and effective performance of targeted models in their long-term retraining process under the disturbance of poisoning attack.

- Compared with K-means defense method in Fig. 2 and Fig. 3, MSCA performs better in defensing against the poisoning samples crafted by BEBP and CBP on NSL-KDD data set. Relatively speaking, MSCA can better defense against the poisoning samples crafted by BEBP, but the experimental results using MSCA and K-means both perform less effective in defensing against the CBP poisoning method.

Table 2: Comparative results of *ACC* on Kyoto 2006+ under 5 round poisoning attack using BEBP

Targeted model	**Defense**	**Round1**	**Round2**	**Round3**	**Round4**	**Round5**
SVM-RBF	No defense	98.64%	97.02%	95.93%	94.06%	92.41%
	MSCA	98.17%	97.42%	96.33%	95.87%	94.72%
	Mean	98.02%	97.13%	95.98%	95.14%	94.24%
	Best	98.36%	97.64%	96.34%	95.81%	95.02%
	Worse	97.31%	96.52%	95.70%	94.32%	93.45%
SVM-linear	No defense	98.67%	97.57%	96.52%	95.18%	93.75%
	MSCA	98.15%	97.69%	96.78%	95.57%	94.38%
	Mean	98.27%	97.60%	96.56%	95.44%	94.20%
	Best	98.48%	97.73%	96.63%	95.63%	94.32%
	Worse	98.15%	97.51%	96.45%	95.09%	94.09%
SVM-Sigmoid	No defense	97.66%	96.08%	94.08%	92.50%	91.03%
	MSCA	95.93%	94.87%	94.41%	93.45%	92.35%
	Mean	95.87%	94.82%	94.39%	93.42%	92.10%
	Best	96.84%	95.35%	94.57%	93.85%	92.36%
	Worse	95.49%	94.35%	93.81%	93.21%	91.47%

Table 3: Comparative results of *ACC* on Kyoto 2006+ under 5 round poisoning attack using CBP

Targeted model	**Defense**	**Round1**	**Round2**	**Round3**	**Round4**	**Round5**
SVM-RBF	No defense	98.57%	96.91%	94.08%	92.05%	90.60%
	MSCA	98.28%	96.69%	95.87%	93.84%	93.21%
	Mean	98.28%	96.43%	95.46%	93.65%	92.89%
	Best	98.54%	96.76%	95.79%	93.93%	93.39%
	Worse	97.81%	95.85%	95.06%	93.14%	92.59%

SVM-linear	No defense	98.39%	97.08%	94.67%	92.78%	91.30%
	MSCA	97.54%	96.48%	95.45%	94.20%	93.23%
	Mean	97.46%	96.37%	95.23%	93.98%	92.97%
	Best	97.87%	96.52%	95.41%	94.26%	93.29%
	Worse	96.67%	96.17%	95.00%	93.44%	92.45%
SVM-Sigmoid	No defense	97.30%	96.34%	94.17%	91.71%	89.93%
	MSCA	97.00%	95.54%	95.06%	94.21%	93.06%
	Mean	96.82%	95.33%	94.66%	93.97%	92.65%
	Best	97.00%	95.49%	94.85%	94.29%	93.21%
	Worse	96.57%	95.23%	94.21%	93.61%	91.96%

- From Tab. 2 and Tab. 3 we can see that MSCA with the average ensemble strategy is more stable than the single spectral clustering method. Although it is difficult to use average ensemble strategy achieving better performance than the best spectral clustering learner, MSCA can achieve the comparable performance with the best one among all base spectral clustering learners. Moreover, MSCA performs significantly better than the worst one as well slightly better than the average *ACC* of all base learners.

6 Conclusions

In this paper, we have proposed a novel sample evaluation method using spectral clustering and ensemble strategy. Firstly, we propose using spectral clustering algorithm to evaluate training samples and rate their *LC* values. To address the drawback of single spectral clustering, we further present the MSCA method to realize more robust evaluation for training sample. Experiments on real intrusion detection data sets demonstrate the proposed sample evaluation method can effectively defense against poisoning sample crafted by different methods.

In future, it is worthwhile to do more in-depth studies on the scalability of the proposed sample evaluation method. Moreover, designing more robust algorithms using the result of sample evaluation will be a worthwhile work as well.

References

Barreno, M.; Nelson, B.; Sears, R.; Joseph, A. D.; Tygar, J. D. (2006): Can machine learning be secure? *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, pp. 16-25.

Biggio, B.; Fumera, G.; Roli, F.; Didaci, L. (2012): Poisoning adaptive biometric systems. *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 417-425.

Biggio, B.; Nelson, B.; Laskov, P. (2012): Poisoning attacks against support vector machines. *Proceedings of the 29th International Conference on Machine Learning*, pp. 1467-1474.

Carlini, N.; Mishra, P.; Vaidya, T.; Zhang, Y.; Sherr, M. et al. (2016): Hidden voice commands. *Proceedings of the 25th USENIX Security Symposium*, pp. 513-530.

Chengzhang, Z.; Longbing, C.; Qiang, L.; Jianping, Y.; Vipin, K. (2018): Heterogeneous metric learning of categorical data with hierarchical couplings. *IEEE Transactions on Knowledge & Data Engineering*, vol. 30, no. 7, pp. 1254-1267.

Dietterich, T. G. (2000): Ensemble methods in machine learning. *Multiple Classifier Systems, First International Workshop*, pp. 1-15.

Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A. et al. (2012): Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97.

Kloft, M.; Laskov, P. (2010): Online anomaly detection under adversarial impact. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 405-412.

Laishram, R.; Phoha, V. V. (2016): Curie: a method for protecting svm classifier from poisoning attack. http://arxiv.org/abs/1606.01584.

Liu, Q.; Li, P.; Zhao, W.; Cai, W.; Yu, S. et al. (2018): A survey on security threats and defensive techniques of machine learning: a data driven view. *IEEE Access*, no. 99, pp. 12103-12117.

Lingyun, X.; Xiaobo, S.; Jiaohua, Q.; Wei, H. (2018): Discrete multi-graph Hashing for large-scale visual search. *Neural Processing Letters*.

Lingyun, X.; Yan, L.; Wei, H.; Peng, Y.; Xiaobo, S. (2018): Reversible natural language watermarking using synonym substitution and arithmetic coding. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 541-559.

Lingyun, X.; Guohan, Z.; Qian, L.; Wei, H.; Feng, L. (2018): TUMK-ELM: a fast unsupervised heterogeneous data learning approach. *IEEE Access*, vol. 6, pp. 35305-35315.

Makili, L.; Vega, J.; Dormido-Canto, S.; Pastor, I.; Murari, A. (2011): Computationally efficient svm multi-class image recognition with confidence measures. *Fusion Engineering Design*, vol. 86, no. 6, pp. 1213-1216.

Nelson, B.; Barreno, M.; Chi, F. J.; Joseph, A. D.; Rubinstein, B. I. P. et al. (2008): Exploiting machine learning to subvert your spam filter. *Proceedings of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats*, pp. 1-7.

Nguyen, A.; Yosinski, J.; Clune, J. (2015): Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427-436.

Pan, L.; Qiang, L.; Wentao, Z.; Dongxu, W.; Siqi, W. (2018): Chronic poisoning against machine learning based IDSs using edge pattern detection. *Proceedings of the 2018 International Conference on Communications*, pp. 1-7.

Pan, L.; Wentao, Z.; Qiang, L.; Xiao, L.; Linyuan, Y. (2018): Poisoning machine learning based wireless IDSs via stealing learning model. *Proceedings of the 13th International Conference on Wireless Algorithms, Systems, and Applications*, pp. 261-273.

Paudice, A.; Munozgonzalez, L.; Lupu, E. C. (2018): Label sanitization against label flipping poisoning attacks. http://arxiv.org/abs/1803.00992.

Rubinstein, B. I.; Nelson, B.; Huang, L.; Joseph, A. D.; Lau, S.-h. et al. (2009): Antidote: Understanding and defending against poisoning of anomaly detectors. *Proceedings of the 9th Internet Measurement Conference*, pp. 1-14.

Song, J.; Takakura, H.; Okabe, Y.; Eto, M.; Inoue, D. et al. (2011): Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation. *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, pp. 29-36.

Tsai, C. F.; Hsu, Y. F.; Lin, C. Y.; Lin, W. Y. (2009): Intrusion detection by machine learning: A review. *Expert Systems with Applications an International Journal*, vol. 36, no. 10, pp. 11994-12000.

Wei, C.; Liang, H.; Longbing, C. (2015): Deep modeling complex couplings within financial markets. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2518-2524.

Wittel, G. L.; Wu, S. F. (2004): On attacking statistical spam filters. *First Conference on Email and Anti-Spam*. http://www.ceas.cc/papers-2004/170.pdf.

Yu, S. X.; Shi, J. (2003): Multiclass spectral clustering. *Proceedings of the 9th IEEE International Conference on Computer Vision*, pp. 313-319.

Zhao, M.; An, B.; Gao, W.; Zhang, T. (2017): Efficient label contamination attacks against black-box learning models. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3945-3951.