




Supplementary Materials: A PSO-SVM for Burst Header Packet Flooding Attacks Detection in Optical Burst Switching Networks

Susu Liu ^{1,†} , Xun Liao ^{1,†}  and Heyuan Shi ^{2,*} 

1. RFECV-RF

In this paper, we choose the RFE with K-fold cross-validation based on random forest (RFECV-RF) for feature selection [1]. Algorithm RFECV-RF consists of two main phases in Algorithm S1.

Algorithm S1: RFECV-RF

Input:
A data with m features $D = \{x_1, x_2 \dots x_m\}$

Output:
The optimal subset of data D_{xi}

```

1 /* Phase 1: RFE */
2 repeat
3   Suppose  $D_x$  is denoted as the feature set, the data after preprocessing  $D$  is
   initialized as  $D_x$ , the  $i$ -th optimal feature subset is  $D_{xi}$ , and its accuracy is set
   as  $AC_{xi}$ ;
4   Establish an RF model, use  $D_x$  for training, and calculate the classification
   accuracy  $AC_{xi}$ . Then calculate the importance measure of the  $i$ -th feature
   using Formula (S3). The calculation result is set to
    $\{Importance_1, Importance_2 \dots Importance_i\}$ , and sort;
5   Exclude the last features of the result sort until  $D_x$  is empty;
6 until the accuracy value  $AC_{xi1}$  of feature subset  $D_{x1}$  is less than  $AC_{xi}$ ;
7 /* Phase 2: Cross-validation */
8 Sort the importance of data features calculated in the Phase 1, and select different
   numbers of feature sets  $\{D_1, D_2 \dots D_n\}$ ;
9 Cross-validate the selected data separately;
10 Find the optimal subset of data  $D_{xi}$ , and the validation is over.
```

1.1. External Estimator: Random Forest

A random forest is constructed as a set of decision trees, each of which contains a set of internal nodes and leaves [2]. The features for internal nodes are chosen based on some criteria, including Gini impurity or information gain for classification tasks and variance for regression. Simply explained, feature importance is the average over the entire forest after counting the 'contribution' made by each feature on each tree. It is evaluated to find the features that are highly correlated with the targeted node state. In this paper, we use the Gini index to evaluate it, which is calculated as

$$GI_a = \sum_{y=0}^4 \sum_{y' \neq k} p_{ya} p_{y'a} = 1 - \sum_{y=0}^4 p_{ya}^2 \quad (S1)$$

where p_{ya} represents the proportion of class y in the node a . The Gini index measures the probability that two samples are randomly selected from node a with different class. The importance of feature x_i at node a , i.e. the Gini index variation before and after the branch of node a , is

$$Importance_{ia}^{(Gini)} = GI_a - GI_l - GI_r \quad (S2)$$

where GI_l and GI_r represents Gini indexes of the two new nodes l and r after branching. Assume that the set B denotes the nodes of feature x occurring in the decision tree b , and that there are B trees in RF. The importance of the feature x_i is

$$Importance_i^{(Gini)} = \sum_{b=1}^B \sum_{a \in A} w_{la}^{(Gini)} \quad (S3)$$

2. Optimization of the SVM by PSO

We introduce the PSO algorithm to adjust the parameters of SVM [3,4]. The structure of PSO is given in Algorithm S2.

Algorithm S2: The structure of PSO algorithm.

```

1 Set the number size of the particle swarm, the initial position and velocity of the
  particles;
2 repeat
3   Use fitness function to evaluate the fitness value of each particle at the current
    position in the population;
4   Get the current optimum of the particle and compare it with pbest;
5   if the current optimum of the particle is higher than pbest then
6     | Update pbest with the current global optimum;
7   end
8   Get the current global optimum and compare it with gbest;
9   if the current global optimum is higher than gbest then
10    | Update gbest with the current global optimum;
11  end
12  Update the calculation of the velocity and position of all particles;
13 until
    the set number of iterations is reached or the optimal fitness value is less than a given threshold;

```

3. Experimental Results

We take the sum of the diagonals as the number of predicted and actual values that are true, and the off-diagonal values as the number of predicted errors. For these four categories, the value of TP is the sum of the numbers on the diagonal, the sum of each column minus the number of TP is added to the value of TN, and the value of the sum of each row minus the number of TP is added to TP. The number of rows and columns in each class is added to the value of TN. Based on the above discussion, TP, TN, FP, and FN of KNN, NB, SVM, and PSO-SVM can be calculated as shown in Figure S1. The first category of PSO-SVM has a TP value of 141, TN of 166, FN of 7, and FP of 9. This indicates that all four categories are basically correctly classified. Among the predicted values of these four models, PSO-SVM has the highest TP and TN, and the lowest FP and FN.

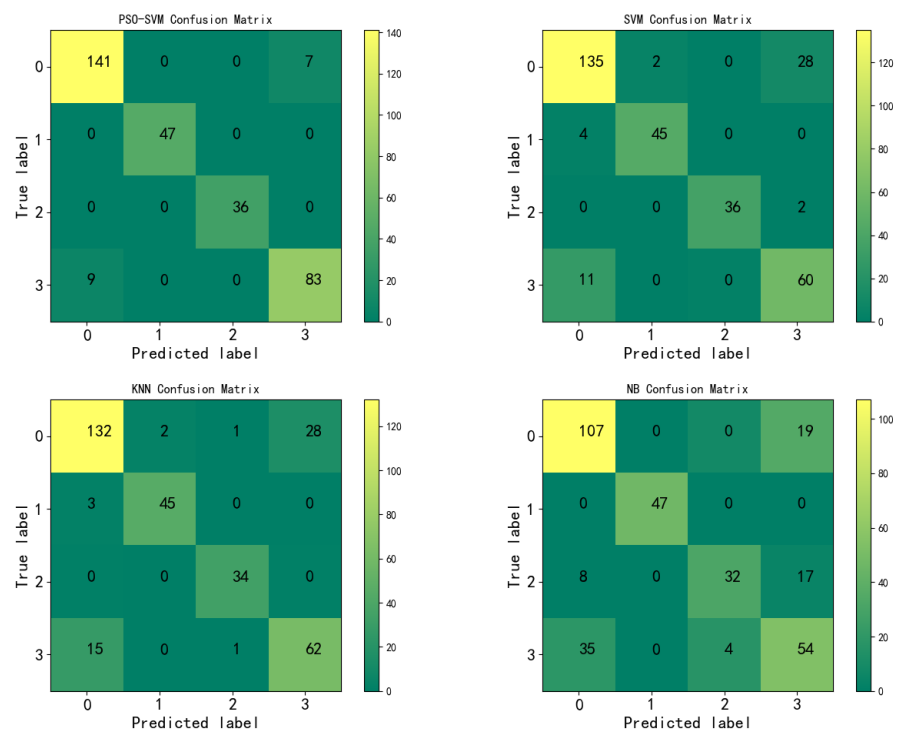


Figure S1. Comparison of the confusion matrix (values) of the four algorithms. 0 ~ 3 indicate the 4 labels, respectively, 0: NB No BLOCK, 1: BLOCK, 2: No BLOCK, and 3: NB WAIT.

References

1. Wang, C.; Pan, Y.; Chen, J.; Ouyang, Y.; Rao, J.; Jiang, Q. Indicator element selection and geochemical anomaly mapping using recursive feature elimination and random forest methods in the Jingdezhen region of Jiangxi Province, South China. *Applied Geochemistry* **2020**, *122*, 104760.
2. Li, X.; Chen, W.; Zhang, Q.; Wu, L. Building auto-encoder intrusion detection system based on random forest feature selection. *Computers & Security* **2020**, *95*, 101851.
3. Hu, J.; Zeng, J.; Wei, L.; Yan, F. Improving the Diagnosis Accuracy of Hydrothermal Aging Degree of V2O5/WO3-TiO2 Catalyst in SCR Control System Using an GS-PSO-SVM Algorithm. *Sustainability* **2017**, *9*, 611.
4. Du, J.; Liu, Y.; Yu, Y.; Yan, W. A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms. *Algorithms* **2017**, *10*, 57.