**Hybrid Machine Learning Model For Network Intrusion Detection**

Group 4 - Term Project Proposal

**References:**

* [A novel SVM-kNN-PSO Ensemble method for network intrusion detection](https://www.sciencedirect.com/science/article/pii/S1568494615006328?casa_token=zmOZ5uTIsxYAAAAA:ORE8DoKGQoBUVqmwOGNAqrVX_kJKyrl5WjXDBsJuAL5KR1UfmJjXOWojLQ4PiicmZPbhtMQ)

**Contributions of Original Paper:**

The original paper uses ensemble-based classifiers to perform the prediction of anomalous and normal connections. They have come up with an anomaly-based detection technique in which they use a group of 6 SVM experts where each expert model consists of 5 individual SVM classifiers for detecting the different categories a connection record can belong to. Similarly, a group of 6 KNN experts is created and these expert systems are combined together to form an ensemble classifier. The paper also uses the original PSO technique, improvised PSO with the LUS method, and WMA voting for weight generation of expert systems. They finally compare the results produced by individual expert systems, and the ensemble classifiers that they implemented using different weight assignment schemes.

**Our results:**

SVM:

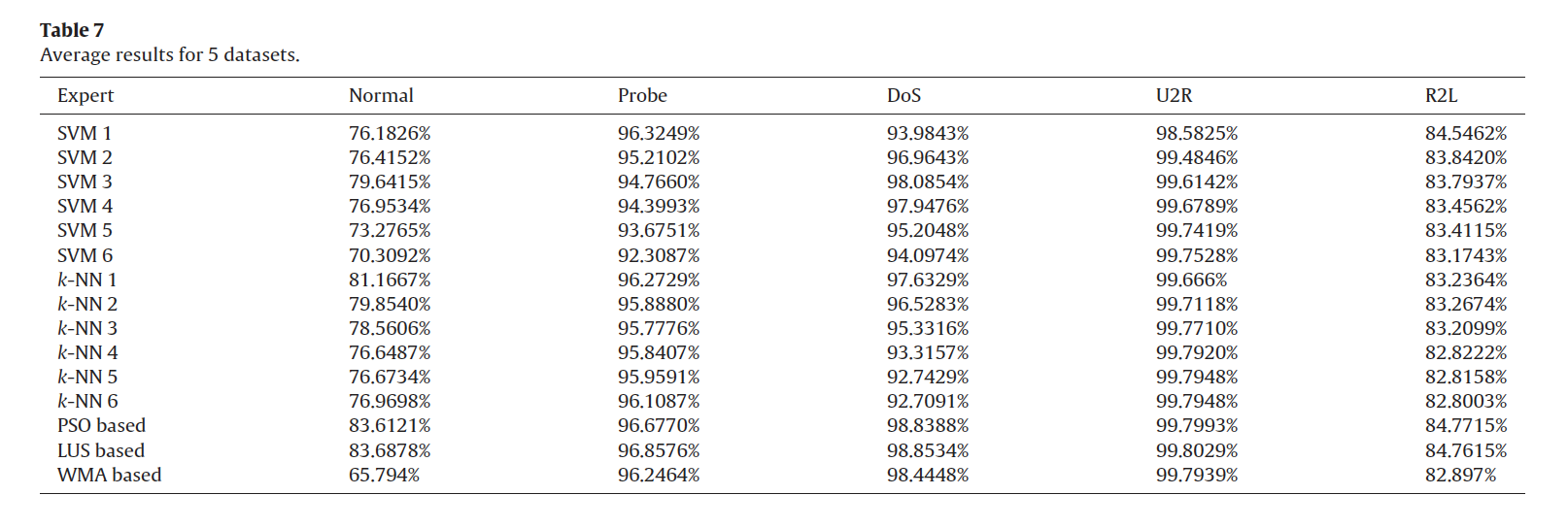
|  |
| --- |
| {'Dos': 91.11022900284462,  'Normal': 85.19155927670067,  'Probe': 92.36619671872188,  'R2L': 94.6703796232644,  'U2R': 99.80472469058732} |

KNN:

|  |
| --- |
| {'Dos': 85.09255876255634,  'Normal': 86.29732813045428,  'Probe': 83.73171883691512,  'R2L': 93.6699043202812,  'U2R': 99.79871543633577} |

PSO Ensemble:

|  |
| --- |
| {'Dos': 90.0,  'Normal': 93.33333333333333,  'Probe': 90.0,  'R2L': 90.0,  'U2R': 90.0} |

**Results mentioned in the Paper:**  

**Problems we ran into :**

Understanding and implementing the concepts of SVM, KNN, and ensemble learning to the content of networks is a great learning experience. During this work, we got to learn about the concepts of network intrusion. What exactly is an intrusion, how it affects the current digital world. The types of intrusions like DoS, U2R, etc., The experience while replicating the work of the authors, it was a difficult thing understanding the preference of certain steps that the authors proceeded. The dataset has a lot of repetitions reducing the quantity and quality of the data. Due to time constraints, we weren’t able to implement the WMA and the LUS optimizers.

**Reasons for the difference in results/failure:**

The expected results and the achieved results are almost similar. The classifier results show slight variation due to the difference in the data attributes selected and data pre-processing. We tried different values of K and RBF functions however we ended up using the values suggested in the paper. We suspect that the randomness of the values being picked in the weight calculation in PSO is responsible for accuracy changes.

**Our Evaluation on the work we tried reproducing:**

As per our evaluation, the KDD99 dataset has given us some extra effort. This is because there are a lot of duplicate instances, unimportant attributes, and missing values. Cleaning the dataset has cost us a lot of time and data loss. Secondly, the values of k for the KNN. We tried different varied ranges of k values for accuracy.

**Comparison of implementations**

**Implementation in Original Paper:**

According to the original paper, their implementation comprises of:

1. Data preprocessing - here they have specifically mentioned the need for converting numerical attributes and deduplication explicitly.
2. Data Classification with 6 different SVM experts: Here the paper explicitly provides us with information about what kernel function to use for SVM - being RBF along with 6 different RBF values to use.
3. Data Classification with 6 different KNN experts: Here the paper explicitly provides us with information of the value for “K” i.e. the number of nearest neighbors to consider for 6 different KNN experts.
4. Data Classification based on ensemble classifier with PSO
5. Data Classification based on ensemble classifier using PSO with LUS improvement
6. Data Classification based on ensemble classifier using WMA

There exist differences in the extent we’re able to implement the steps performed in the original paper. Though we did complete the data-preprocessing, SVM, KNN, and simple PSO parts as suggested in the paper, we missed implementing the further steps of PSO with LUS base ensemble classifier and WMA based ensemble classifier.

**Our Implementation Explained in Detail:**

There are four main steps in this process.

* Data Preprocessing:

The major steps involved here are deduplication, converting non-numerical attributes to numerical attributes, feature selection, and one hot coding of target class variables.

* SVM

SVM(support vector machines) method is used to classify different attack types. We still used the One-Versus-All approach and the RBF kernel function. And we trained six different SVM experts with different RBF parameters, which are [5 2 1 0.5 0.2 0.1] same as the original paper.

We imported several functions from the sklearn package, especially the OneVsRestClassifier function for the One-Versus-All approach. Then training and predicting for the model. And calculating the average accuracy for each attack type, we can detect those attack types with high accuracy.

* KNN

The K - nearest neighbor algorithm, the reason why this algorithm is included in the ensemble procedure is that it works with similar experts as the SVM. The functionality of KNN is as follows:

The KNN algorithm is based on the selected k value, there is a k number of pools. When a new point appears, based on the euclidean distance metric, its distance with the pools is calculated. To validate these weights, the 5 different experts vote their opinion. If the expert believes that the point belongs to that pool, it gives a 1 else 0. Based on the voting count, the point is classified to the pool or it passes.

So, for this procedure, as earlier, we divide the data into 6 different parts. The model is trained for these 6 datasets. The test data is used to determine the accuracy of the model. For these 6 sets, the average prediction score is calculated for the different types of network intrusions. Using sklearn’s cross-validation we determine the accuracy of the model.

The accuracy depends on the selected k value. Varying it may improve or decrease the accuracy. So for a range of k values, the model is trained and tested for the best accuracy.

Accuracy of the different network intrusions using knn are as follows:

* Ensemble PSO

Let us begin understanding the meaning of ensemble. It is the process of merging different classification algorithms to form a complex and comprehensive classifier to improve accuracy and performance.

Now the next step is to know, why do we need a PSO. The PSO called Particle Swarm Optimizer which is a methodology that helps in adding weights to the selected experts which reduces the problem of optimization. In this algorithm, there are two important calculations. One is velocity and the other is position. Initially, a set of random positions are taken. Using the fitness function and the velocity the correct position of these points is determined.

There are other optimization techniques like LUS (Local Unimodal Sampling), WMA (Weighted Majority Algorithm). The selection of PSO has specific reasons. One is that this is a derivative-free, zero-order method.

Secondly, the primal attributes of any algorithm are Speed and Accuracy. The LUS algorithm is on a similar scale with respect to the accuracy of the PSO. But the speed is very slow. The LUS takes nearly 500 times more time than the PSO. Whereas WMA is very fast but has a big penalty on its accuracy.

.