Weekly Report of Week 2

Name: Punyawat Jaroensiripong

Topic: Model Selection and Training

Description: In last week, I analyzed the dataset called “NSL-KDD” dataset which contain the traffic log of the multiple attack. From last meeting with professor, I have been assigned the topic to find the state of the art of this dataset. Therefore, I research into the paper and choosing the well-known Machine learning model as follows:

1. Logistic Regression
2. GaussianNB (Gaussian Naive Bayes)
3. Decision Tree Classifier
4. Random Forest Classifier
5. C-Support Vector Classification (SVC)
6. KNeighborsClassifier (KNN)
7. Gradient Boosting Classifier
8. Multi-layer Perceptron classifier (MLPClassifier)
9. Ada Boost Classifier
10. XGBoost Classifier

In the preprocessing and feature selection, I used One-hot-vector to change the string data from the dataset feature which is 'protocol_type', 'service' and 'flag' and assign the string feature into integet type in binary form (0, 1) to the new columns. At first, the model training without feature selection (Did not cut off any features) has shown the high accuracy. But when plotting the learning curve, the graph has shown the overfitting problem that happened in the model. Therefore, feature selection is one of the solutions that will reduce the overfitting problem. Thus, I split the preprocessing into multiple way, as follows:

1. Normal Dataset
2. Feature Selection from last week analysis
3. Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction and data exploration. It identifies the most significant patterns and relationships in high-dimensional data by transforming it into a lower-dimensional representation. And using logarithmic scaling to scale the dataset.
4. Genetic algorithms (GA) are optimization techniques inspired by the process of natural selection and genetic inheritance. They mimic the principles of evolution to search for optimal solutions in complex problem spaces.

In the training section, I split the ‘KDDtrain+’ into training set with 80% and test set 20% and using the test set as mentioned and ‘KDDtest+’ to test the accuracy.

The result show that the Normal dataset, feature section from last week analysis and GA have high accuracy score but when analyzing the learning curve, the graph has shown the overfitting. On the other hand, the PCA show to low accuracy score, but the learning curve shown the acceptable result learning. Hence, in this week I still cannot make a conclusion on the Machine learning model that will be fitted in to the dataset.

The obstacle in this week was the finding the right method to prevent and improve the model that will not be overfitting.

On the future work in next week, I'm planning to work on implementation deep learning model from the paper that provide the details of each dense of LSRM-RNN model. Moreover, I have an idea to transform the traffic log data which contain attack into image data before analyzed in machine learning model or deep learning model.