SIT719 Security and Privacy Issues in Analytics

Credit Task 4.2: Intrusion Detection using Supervised Learning Techniques

Overview

An Intrusion Detection System (IDS) is a system that monitors network traffic for suspicious activity and issues alerts when such activity is discovered. Supervised learning techniques have been proven very effective for intrusion detection.

This week, we have explored the supervised machine learnings and how to apply them for security solutions, e.g., intrusion detection in NSL-KDD dataset. We have demonstrated how classification algorithms can separate the normal instances from the attack classes. To perform supervised classification tasks, we have employed WEKA.

A basic tutorial for the new users can be obtained from the following link:
https://www.youtube.com/watch?time_continue=198&v=TF1yh5PKaqI&feature=emb_logo

If you want to learn more on how WEKA can be used, please follow the below reference:
https://waikato.github.io/weka-wiki/documentation/

This is a Credit task, so please make sure you are already up to date with all previous ***Pass tasks*** before attempting this task.

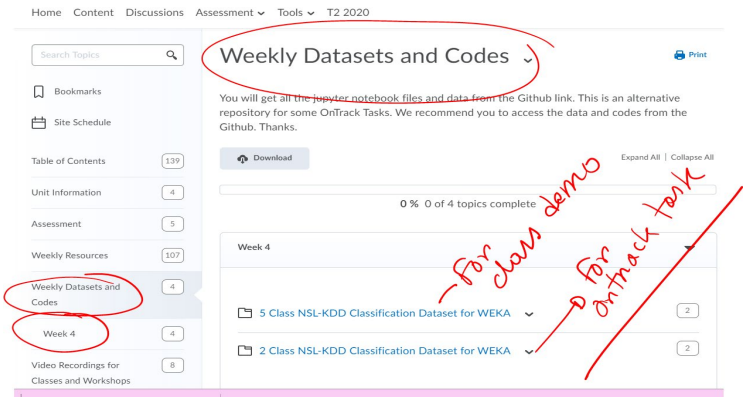
Note: This is an extension of the *pass task* of 4.1P. Therefore, you can use the results obtained in *pass task 4.1* to answer the initial 5 steps of this task during your report preparation. If you are attempting this one, please make sure you also submit 4.1P separately.

Task Description

Instructions:
This task is a binary classification. Follow the below steps to complete the task. Once you have the results and reports, compile in a PDF and submit to the onTrack system.

This task is a binary classification (2-class problem). Follow the below steps to complete the task. Once you have the results and reports, compile in a PDF and submit to the onTrack system.

1. Download the data folder from the CloudDeakin contents, look for Weekly Dataset and codes. Then look for 2-class NSL-KDD datasets (both train and test) and save in a folder.



Load the Train dataset into WEKA. Once uploaded, you may check the data distribution by selecting the class attribute and it will appear as Figure 2.

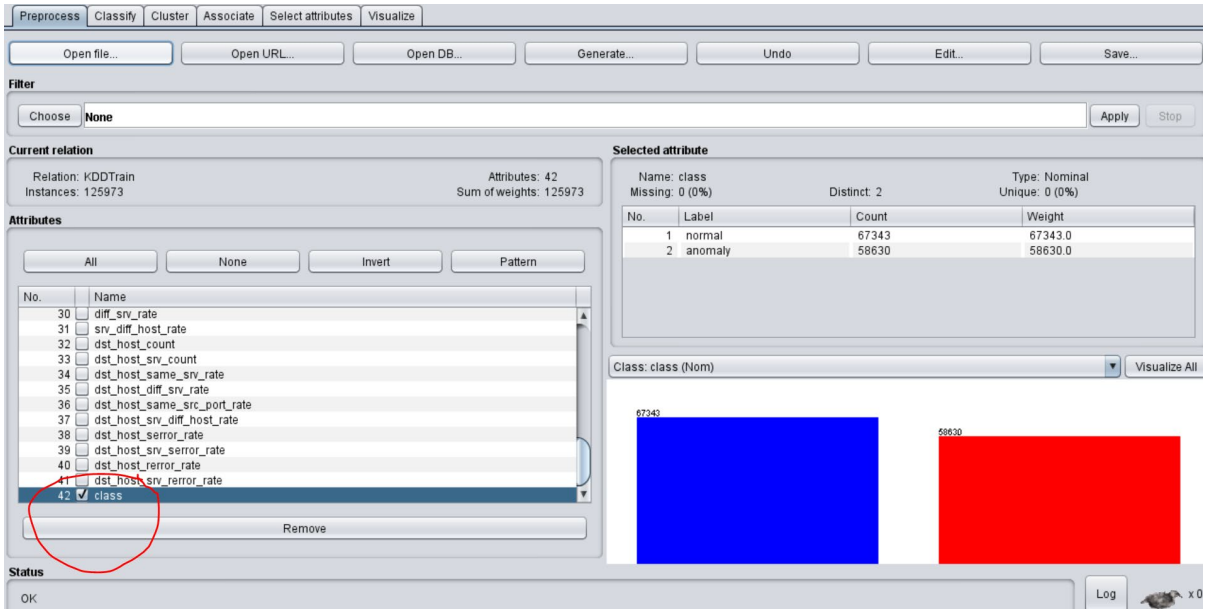


Figure 2

2. Now apply “Naïve Bayes” classification algorithm from the “Classify” tab.

3. Check the results with a 10-fold cross validation.

4. Now, upload the test dataset and check the classification results.

5. Compare the results between 10-fold cross validation and the one obtained using the test dataset. Use confusion matrix to explain the results. Also, include a brief description on 10-fold cross validation.

The above 5-steps are similar to the pass task for this week. Therefore you can use the results from there while preparing this report. Address carefully the below questions to obtain a *credit grade.

6. Similar to the "Naïve Bayes", apply ***at least 5 other supervised classification techniques*** and compare their performance. To report the performance create a table and present the following measures. Then compare the outcome of your nominated 5 algorithms. You can choose any 5. However try to consider high performing algorithms.

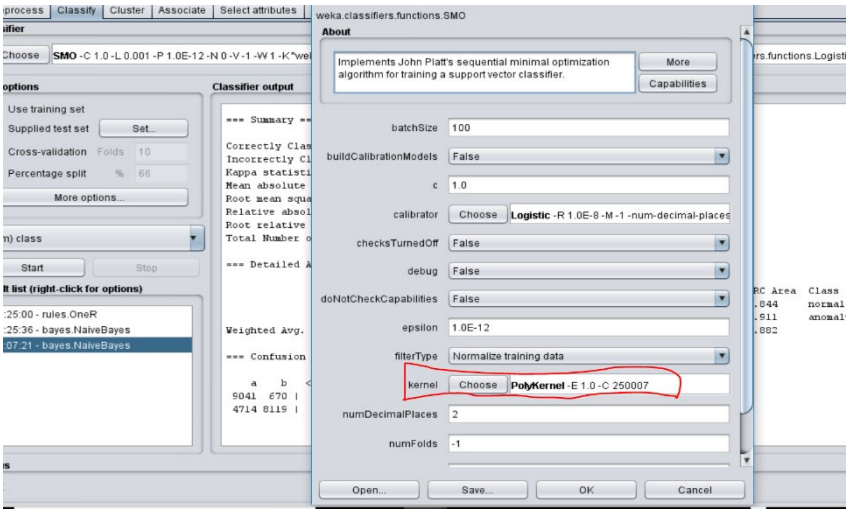
- TP Rate
- FP Rate
- Precision
- Recall
- F-Measure
- ROC Area

Sample Table:

Algorithms	TP Rate	FP Rate	…	…	…	…	...
1.							
2.							

Note: It is a large dataset. Therefore, some classifier may take huge amount of time if the machine processing power and storage is not too high. In such cases, try to avoid those.

7. Some algorithms may have tuning parameter. Consider the SMO based SVM algorithm. You can try different kernel trick as shown below. Change the kernels to "PolyKernel" and ensure that the filter normalizes the training data as shown in the figure. If you start the task, it will take too much time on this large dataset. So you need to reduce the sample size of the dataset to make it manageable (note: it may impact on the performance).



Therefore, you need to resample and make the data size to 20% of its original size.

Hint: To do that go to the choose option in the preprocess tab, and then go to filters>supervised>instances>resample and choose the right parameters.

8. Now perform classification task based on SVM classifier (SMO) using POLY and RBF kernels and report the confusion matrices and computation time.

Finally, produce a minimum 500-word report summarizing all results and submit the compiled PDF to the onTrack system.

Please note, it is a graded task where you will receive some feedback and marks. Your tutor/marker will assign you some marks based on the quality of your submission, how well you have explained the results, your usage of scientific language, authenticity of the claims and finally the aesthetic look of your submission and reflection of the quality of your work from the tutor's judgement.