ISTANBUL TECHNICAL UNIVERSITY
FACULTY OF COMPUTER AND INFORMATICS ENGINEERING

Project Description Form for Computer Project II Course

PROJECT 1B

Assignment Date: 21.02.16

Submission Date: 15.03.2016 23:00

Instructor(s): Zehra Çataltepe (ITU, tazi.io), Tanju Çataltepe (@tazi.io)

Project Name: Machine Learning for Network Intrusion Detection

Project Duration: 4 weeks

Project Workload (man-hour): 40 hour/man

Project Weekly Plan:

1. Software installation and dataset preprocessing
    a. Install Spark MLLib on your computer. Identify why you would need Spark MLLib as opposed to other machine learning solutions such as weka. Test that the decision tree algorithm works using a simple dataset that will be provided.
    b. Get the KDD99 Network Security (10 percent) data (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html) from http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gz.
    c. Describe what the attacks in the dataset mean? How were the features in the dataset produced?
    d. Partition the data into a training and validation set in such a way that for each of the 23 classes the last 10% is in the validation set and the first 90% is in the training set.
    e. Prepare three datasets where the classes are identified as D23: 23 classes (22 attacks and 1 normal), D5 :5 classes (4 attacks: probe, dos, u2r, url, and normal), D2: 2 classes (attack, normal).
    f. Prepare another three set of unbalanced datasets (name them as U23 (from D23), U5 (from D5), U2 (from D2)) where you keep an instance if it is normal. If the instance is an attack, you keep it in the dataset with only 10% probability.
2. Spark MLLib Decision tree performance: Train a classifier on training dataset and test it on the validation set. Measure the training and testing time, classification accuracy, confusion matrix for the D23, D12 and D2 datasets. Determine the best parameters of the decision tree for the best accuracy.
3. Performance improvement for unbalanced data: Measure the training and testing time, classification accuracy, confusion matrix for the U23, U12 and U2 datasets.
4. Compare the training and validation performances of the MLLib for the six different datasets. Use a browser based decision tree visualization tool to visualize your decision trees. Compare the decision trees produced.

Keywords: Machine Learning, Intrusion Detection, Decision tree, Spark MLLib, Storm Samoa

Success Criteria: Students will implement decision tree models which has more than 80% classification accuracy on the KDD99 dataset.

Tools:

1. Spark MLLib as ML tool
2. Java/Scala for data preprocessing and performance comparison.

Grading:

1. Spark MLLib Installation & Usage : %5
2. Data preprocessing : %20
3. Spark MLLib training and validation performance original datasets D23, D5 and D2: %25
4. Spark MLLib Performance on unbalanced datasets U23, U5, U2: %25
5. Results interpretation and presentation: %25

References: https://spark.apache.org,

Project Description: In this project, students will experiment with the Spark MLLib platform and the network intrusion detection dataset KDD99. Decision tree will be used as the classifier. Students will gain experience on performance of the classifiers for different types of datasets, for example when the class description changes or the classes are unbalanced.