Summary

by Ibrahim Ahmethan

Submission date: 06-May-2022 03:46PM (UTC+0100)

Submission ID: 178827270

File name: Summary.pdf (112.63K)

Word count: 513

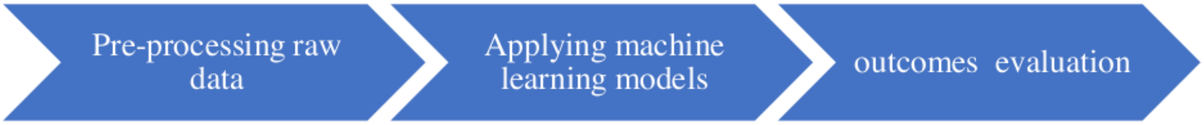
Character count: 3959

Overview

Variety of machine learning models and data pre-processing techniques was implemented in KDD-CUP-99 to achieve and the highest possible accuracies and outcomes, two main approaches were followed for this problem by using Python programming language on Jupyter Notebook platform.

1- data pre-processing (to be used later in the machine learning models)
2- applying supervised machine learning models

however different data processing techniques and machine learning models achieved different results and outcomes, the following diagrams showing the summary of the task.



Pre-processing raw data

- data cleaning
- feature engineering

Applying machine learning models

- binary attack class classifications
- multi-class attack classifications

outcomes evaluation

- accuray testing
- metrics evaluation
- ROC and epoch vs accuracy graph

Pre-processing raw data

To pre-process the raw data, data cleaning process applied by removing duplicated rows, detecting outliers and checking for missing values then feature engineering techniques applied such as feature encoding (to convert categorical data into numerical data), feature scaling (to scale down the range of numerical data) and feature selection (to reduce the dimension and complexity of the dataset), in summary the following approaches implemented in the stage of feature engineering

1- OneHotEncoder to encode categorical values
2- Standardization to scale down the numerical values
3- principal component analysis (PCA) and correlation to to reduce the dimensionality and complexity of the dataset
4- visualizing and analyzing the dataset to ensure that the dataset is ready to be used in ML models

Applying machine learning models

Three types of supervised machine learning model implemented, MLP, KNN and Quadratic Discriminant Analysis Classifier, most of them achieved accuracy higher than 97 % and showed success in different evaluation metrics and graphs.

Conclusion

As a result of well organized and pre-processed dataset, carefully designed topologies of ML used and variety of evaluation techniques applied, the system succeeded to show high performance in both binary and multiclass classification task.

❖ References

- https://www.researchgate.net/publication/48446353_A_detailed_analysis_of_the_KDD_CUP_99_data_set

- http://proceedings.mlr.press/v7/niculescu09/niculescu09.pdf

- https://www.sciencedirect.com/science/article/pii/S0925231219315759

- https://www.researchgate.net/publication/330722518_Intensive_Pre-Processing_of_KDD_Cup_99_for_Network_Intrusion_Classification_Using_Machine_Learning_Techniques

- https://towardsdatascience.com/need-for-feature-engineering-in-machine-learning-897df2ed00e6

- https://medium.com/analytics-vidhya/building-an-intrusion-detection-model-using-kdd-cup99-dataset-fb4cba4189ed

- https://www.youtube.com/watch?v=UV2tawxGhyw

- https://vitalflux.com/machine-learning-feature-selection-feature-extraction/

- https://elitedatascience.com/dimensionality-reduction-algorithms#:~:text=4.,extraction%20creates%20brand%20new%20ones.

- https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

- https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/

- https://www.youtube.com/watch?v=MPnNC6kkNC4

- https://www.youtube.com/watch?v=6WDFfaYtN6s

- https://medium.com/geekculture/network-intrusion-detection-using-deep-learning-bcc91e9b999d

- https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af

- https://www.analyticsvidhya.com/blog/2021/10/implementing-artificial-neural-networkclassification-in-python-from-scratch

- https://stackoverflow.com/questions/37213388/keras-accuracy-does-not-change

- https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network/#:~:text=Therefore%2C%20an%20MLP%20that%20has,of%20nodes%20in%20each%20layer.

- https://scikit-learn.org/stable/

- https://pandas.pydata.org/

- https://keras.io/

- https://numpy.org/

- https://www.tensorflow.org/resources/learn-ml?gclid=CjwKCAjwjtOTBhAvEiwASG4bCGLTHNuJLlDlt_nkz_YzqaOt6oahLzy_-WRjmHrGOZhfTFoZjFv0JxoCX4IQAvD_BwE

- https://matplotlib.org/

- https://seaborn.pydata.org/

- https://learn.g2.com/data-preprocessing

- https://serokell.io/blog/data-preprocessing

- https://towardsdatascience.com/all-machine-learning-models-explained-in-6-minutes-9fe30ff6776a