Kddcup 99 dataset

The KDD Cup '99 dataset was created by processing the tcpdump portions of the 1998 DARPA Intrusion Detection System (IDS) Evaluation dataset, created by MIT Lincoln Lab [2]. The artificial data (described on the dataset's homepage) was generated using a closed network and hand-injected attacks to produce a large number of different types of attack with normal activity in the background. As the initial goal was to produce a large training set for supervised learning algorithms, there is a large proportion (80.1%) of abnormal data which is unrealistic in real world, and inappropriate for unsupervised anomaly detection which aims at detecting 'abnormal' data, i.e.:

- qualitatively different from normal data
- in large minority among the observations.

We thus transform the KDD Data set into two different data sets: SA and SF.

- SA is obtained by simply selecting all the normal data, and a small proportion of abnormal data to gives an anomaly proportion of 1%.
- SF is obtained as in [3] by simply picking up the data whose attribute logged_in is positive, thus focusing on the intrusion attack, which gives a proportion of 0.3% of attack.
- http and smtp are two subsets of SF corresponding with third feature equal to 'http' (resp. to 'smtp').

General KDD structure :

Samples total	4898431
Dimensionality	41
Features	discrete (int) or continuous (float)
Targets	str, 'normal.' or name of the anomaly type

SA structure :

Samples total	976158
Dimensionality	41
Features	discrete (int) or continuous (float)
Targets	str, 'normal.' or name of the anomaly type

SF structure :

Samples total	699691
Dimensionality	4
Features	discrete (int) or continuous (float)
Targets	str, 'normal.' or name of the anomaly type

http structure :

Samples total	619052
Dimensionality	3
Features	discrete (int) or continuous (float)
Targets	str, 'normal.' or name of the anomaly type

smtp structure :

Samples total	95373
Dimensionality	3
Features	discrete (int) or continuous (float)
Targets	str, 'normal.' or name of the anomaly type

:func:`sklearn.datasets.fetch_kddcup99` will load the kddcup99 dataset; it returns a dictionary-like object with the feature matrix in the `data` member and the target values in `target`. The "as_frame" optional argument converts `data` into a pandas DataFrame and `target` into a pandas Series. The dataset will be downloaded from the web if necessary.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\sklearn\datasets\descr\[scikit-learn-main][sklearn][datasets][descr]kddcup99.rst`, **line 78**); *backlink*

Unknown interpreted text role "func".

References

[2] Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation, Richard Lippmann, Joshua W. Haines, David J. Fried, Jonathan Korba, Kumar Das.

[3] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne. Online unsupervised outlier detection using finite mixtures with discounting learning algorithms. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 320-324. ACM Press, 2000.