**Abstract**

Dataset pruning is the process of removing sub-optimal tuples from a dataset to improve the learning of a machine learning model. In this paper, we compared the performance of different algorithms, first on an unpruned dataset and then on an iteratively pruned dataset. The goal was to understand whether an algorithm (say A) on an unpruned dataset performs better than another algorithm (say B), would algorithm B perform better on the pruned data or vice-versa. The dataset chosen for our analysis is a subset of the largest movie database publicly available on the internet, IMDb [1]. The learning objective of the model was to predict the categorical rating of a movie among 5 bins: poor, average, good, very good, excellent. The results indicated that an algorithm that performed better on an unpruned dataset also performed better on a pruned dataset.

**Introduction**

A fine line separates cleaning and pruning of a dataset. Cleaning mostly is a preprocessing step [2] that involves removing unrequired data, data imputation, standardizing or normalizing the feature ranges and converting categorical values to numbers [3]. In comparison pruning takes place after preprocessing, where certain data is strategically removed to improve the machine learning model. In this paper we try to bring forth the effect of dataset pruning on the performance of different machine learning algorithms, i.e. If an algorithm (say A) on an unpruned dataset performs better than another algorithm (say B), will algorithm B perform better on the pruned data or vice-versa? This interest has been taken forward as we trained our model using classification algorithms – Logistic Regression, Support Vector Classifier (SVC) and Random Forest Classifier, initially on the unpruned dataset and then iteratively pruned the dataset and compared the evaluation metrics of each iteration. The evaluation metrics used in this paper are Accuracy score and F1 score.

[1] https://link.springer.com/content/pdf/10.1007%2F978-3-319-11740-9\_41.pdf

[2] <https://towardsdatascience.com/data-cleaning-101-948d22a92e4>

[3] <https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python/>

---

In the past, there have been a lot of research activities on various predictive algorithms, followed by its measure on accuracy and consequently its comparison. It is important to have an unbiased dataset as it will help the researchers to test their hypothesis in an able manner. Most of the raw datasets require cleaning i.e. removal or correction of corrupted data. After cleansing, pruning the cleaned data is required to make accurate predictions. The technique also reduces the risk of overfitting or under-fitting. As a part of this machine learning group project, we tried to find out how much pruning the dataset actually affected the performance of learning algorithms. To elaborate, if pruning can be applied on a cleansed dataset by removing the sub-optimal features and then checking the accuracy. If an algorithm (say A) on an unpruned dataset performs better than another algorithm (say B), will algorithm B perform better on the pruned data or vice-versa? This interest has been taken forward as we picked an uncleaned movie dataset from Kaggle, applied pruning techniques and ran algorithms on the pruned dataset.

---