Data Mining I (1DL360)
Uppsala University – Autumn 2019
Report for Project

Group 21 - Venkata Sai Teja Mogillapalle, Dhanush Kumar Akunuri, Khan Asif

October 27, 2019

1 Introduction

With the increase in the data that's being generated everyday (due to technological advancement), the necessity to store large amounts of data has increased. Many data centers have come up to encourage the users to store their data online(cloud). The data centers are equipped with thousands of hard drives to store the data and thus preventing the data loss (in case of failure of the hard drive) is a very important factor for the data centers. Since there would be thousands of hard drives, storing the back-up data of all the hard drives would involve a lot of cost and resource. One way to solve this issue is to predict the failure of the hard drive before they occur so that they can just back-up the potential hard drives that are prone to fail.

Backblaze is a data storage provider and they provide the daily log information of the hard drives in their data centers. We use this data to predict the failures of the hard drives. We will use Classification task to classify a hard drive into 2 classes: Failure or Not Failure.

2 Data

Our dataset contains basic hard drive information and there are almost 90 columns of raw and normalized S.M.A.R.T(Self-Monitoring, Analysis and Reporting Technology) values. We collected 2016 Quarter 1 data from Backblaze website. Basic information in the data includes

- **date**: yyyy-mm-dd format (Ordinal)
- **serial_number**: Unique identifier for hard drive. (Nominal)
- **model**: Manufacturer-assigned model number of drive. (Nominal)
- **capacity_bytes**: Drive capacity in bytes. (Ordinal)
- **failure**: '1' if the hard drive fails and '0' if it doesn't fail.(Nominal)
- **S.M.A.R.T**: All the S.M.A.R.T values are of numeric data type. (Interval)

3 Preprocessing

There are a total of 65993 different hard drives in the data. An entry for each hard drive is made every day until the hard drive fails. For the first quarter of 2016, a total of 3179295 entries have been made. But for around 66000 hard drives , there were only 215 failures observed in the data. So, we could see a lot of imbalance in the data (Number of failures to total number of hard drives present) and also the SMART stats can vary in meaning based on hard drive manufacturer and model. So, we filtered the data based on the "model" and took the model which has most number of failures. For model "ST4000DM000", there were 139 failures and a total of 35057 different hard drives for that model.

Next, we removed the rows which has "NULL" attributes. After that, we checked for the columns which has a constant value and removed that column as it doesn't show any effect on the prediction.

Since we have over 90 columns in the data, it becomes really difficult to apply any algorithm and classify the data. Backblaze, the data provider, has pointed out some columns which have high correlation for hard drive failures. So, we applied Dimensionality reduction to the data and took 21 columns for S.M.A.R.T attributes.

4 Results

4.A Set-up

Since we have a classification task here(Failure or Not Failure), we chose Random forest algorithm to classify our data. We split the data into training data and test data. 75% data is used for training and 25% data is used for testing. Random forest is trained with the training data and the resulting model is used for testing.

4.B Evaluation/Discussion

Our Random forest algorithm predicted a total of 17 failures in the test data. Out of which 15 failures have actually shown up in the data. The test data has a total of 31 failures. After applying the algorithm, the following perfomance measures have been recorded:

- **#True Positives** : 15
- **#False Negatives** : 16
- **#False Positives** : 2
- **#True Negatives** : 8731
- **Precision** : 0.882
- **Recall** : 0.48
- **logloss**: 0.032587668146748634
- **ROC_accuracy**: 0.8840105938542347

Since we have filtered out the data based on the ”model” and we trained our Random forest with only hard drives of the model ”ST4000DM000” , the accuracy of the model has been improved greatly. Our Random Forest identified that 17 hard drives would fail and out of them 15 have reported as failures. Our model has recorded a very low false-alarm rate which is a good thing.

5 Conclusion

By applying the Random Forest algorithm, we were able to predict the failures of the hard drives based on their daily recorded S.M.A.R.T values. This prediction of failures is a very crucial information for data centers as they could find the hard drives which might fail in the future and they could take a back-up of the those hard drives in advance which will prevent the data loss.