MACHINE LEARNING APPROACHES TO HANDLE "BIG DATA" FROM PAMS DEVICES

Batishahe Selimi[1], Arvind Thiruvengadam[1], Saroj Pradhan[1]
bs0035@mail.wvu.edu, arvind.thiruvengadam@mail.wvu.edu,
saroj.pradhan@mail.wvu.edu

[1] West Virginia University

Portable Activity Monitoring Systems (PAMS) are increasingly becoming a versatile tool to understand the day-to-day operation of vehicles and their impact on the environment. PAMS can transmit a large amount of data retrieved from both vehicle ECU communication and external sensors fitted to the vehicle. The increase in the amount of data collected from PAMS opens up huge possibilities, but it comes with complexities in data analysis. We need to be able to store, manage and process large quantities of data in quasi-real-time for efficient use of PAMS information. So, big data is not just large volume of data, but is also the multitude of different sources, formats and data rate with which the data has been transferred.

This presentation will focus on some of the approaches used at West Virginia University (WVU CAFEE) to manage large data transmitted by PAMS. We use data mining and machine learning algorithms to gain knowledge from the data. An important step when dealing with PAMS as any real-world data, is pre-processing. This is necessary due to identify incomplete data (e.g. missing ECU channel of interest), noisy data (containing errors/default values or outliers), and inconsistent data (same ECU channel named differently in different controller or broadcasted in different SPN number).

This work focuses on identifying the thresholds used to sanitize the data to segregate the useful data from the routine communications of the PAMS. A robust quality check of the data is essential for the data analysis part of the process. Algorithms such as unsupervised machine learning (ML) require quality data to predict and disseminate useful statistics from the activity. However, segregation of good data from large dataset is a complex process because an in-depth knowledge of real-world activity is required to distinguish between faulty data and real-world scenarios. This presentation will detail the use of open-source python powerful tools such Pandas, TensorFlow and Keras. Using these frameworks, large scale data transformation and analysis are performed in order to develop state of the art machine learning algorithms.

Finally, the presentation will detail the machine learning approaches currently being researched at WVU CAFEE. The ML approaches researched at CAFEE are aimed towards discerning deteriorating trends in emissions, fuel consumption and vehicle health. ML approaches are either designed for predicting vehicle activity or predicting the onset of failure by closely monitoring chosen engine parameters. Various levels of fault diagnostics can be overplayed over the ML algorithm to perform an offline On-Board Prognosis (OBP) and/or On-Board Diagnosis (OBD) of complex systems such as a heavy-duty vehicle.