**Assignment 2 - Data Scavenger Hunt**

The given dataset consists of data such as Linear Acceleration (LA), Rotational velocity (RV) and Rotational acceleration (RA) and each comprises of 65 instances. The data must be separated so that we can clearly distinguish all the instances upon viewing their attributes. It is difficult to obtain solutions when the data is not well separated since the graphs can overlap each other. Graphs such as Line plots, scatter plots and histograms are often used to visualize data and analyze it. For the given dataset consisting of 65 instances, plots for instance 56 and instance 8 are shown below:

Chart

Description automatically generated

(Fig 1)

Chart

Description automatically generated with medium confidence

(Fig 2)

Upon viewing the plot for the instance 56 (Fig 1), the LA, RA and RV tend to be quite similar to each other showing a steep rise at the beginning. On the other hand, the plot for instance 8 (FIG 2) shows different fluctuations and are different compared to one another.

**Feature Selection:**

For each individual waveform, the min, max and the average has been calculated to identify the features out of the given 65 instances. Feature selection is done to obtain those features that will contribute the most to finding a solution to the problem. Individual graphs have been generated such as MLA (minimum linear acceleration), MRA (minimum rotation acceleration) and so on. Upon viewing these 9 plots, it is clear that each of them consists of a different range and therefore, we make a decision by comparing their ranges. Hence, PLA, PRV and PRA have been selected.

|  |  |
| --- | --- |
| **Feature** | **Instance Numbers of Top 5 largest feature values** |
| PLA | 8,9,11,19,20 |
| PRA | 8,9,11,19,20 |
| PRV | 8,9,11,33,38 |

PLA – Peak Linear acceleration, PRV – Peak Rotational Velocity and PRA – Peak Rotational Acceleration have been selected as the top features since, the max values are computed and can be clearly shown in the plots for each of its instances. Since we are dealing with waveforms, the peak value is the maximum value attained during one cycle. On the other hand, features such as MLA, MRA and MRV show 0 since it’s calculating only the minimum values and does not provide much information. Similarly, ALA, ARA and ARV computes the average. In this case it would fail to provide accurate information when the data is limited. If symmetrical waveforms are to be considered, then the average of both the positive and negative half cycles would be zero. Hence, these features can also be ignored. Therefore, PLA, PRA and PRV can be selected. With the help of a scatter plot, PLA and PRA have the largest feature values for instances: 8,9,11,19,20. Hence, they are capable of selecting the 5 from 65 data instances. A picture containing histogram

Description automatically generated

Company name

Description automatically generated