Knowledge Mining and Big Data Part 3

Pedram Ghazi

Student Number: 267640

1) A) Total number of lines in one day: 131*24*60*60

B) Real sample rate for all buses could be calculated by dividing the total number of samples by the duration: 50,000/303sec = 165.016
Then for each bus we should divide the answer by number of buses:
165016/131 = 1.26 per second is the sample rate for each bus.
It took about 5 minutes and 3 seconds which is 303 seconds totally.
131 vehicles are included.

C) Min = -268
Max = 3594
StdDev = 404.787
Bus with the maximum value: 99

D) An observation point would be called an outlier if it is distant from other observation points. An outlier point occurs because of variability in the measurements or experimental errors. It is up to us to define a method in which the distance of a point indicates if it is an outlier or not. In here I would use a method in which I eliminate the points with value lower than [Q1 – 1.5*D] and with values higher than [Q3+1.5*D]. In these equations "Q1", "Q3", and "D" are respectively, 1st quartile value, 3rd quartile value, and Distance between Q1 and Q3.
Q1 – 1.5*D = 0 – 1.5*147 = -220.5 Then the threshold for the lower band is -220.5
Q3 +1.5*D = 147 + 1.5*147 = 367.5 Then the threshold for the higher band is 367.5

E) The "DestinationName" and "OriginName" attributes in some of the rows in the dataset are empty and we can filter these empty values and change them to a constant value. For "OperatorRef" attribute we can see that the values are entered in different shapes in terms of lowercase and uppercase, for example we have both values "TKL" and "tkl" as "OperatorRef" but this does not make sense since we mean one company by both of those terms. So, these problems also should be corrected.

F) The only unusable attributes were with type string, so after removing them 16 attributes were left.

2) A) Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.
The method also evaluates significance by searching the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility

B) Selected attributes: 9,10,11,12,13 : 5

DirectionRef
DatedVehicleJourneyRef
OperatorRef
VehicleLocationLongitude
VehicleLocationLatitude

It seems these 5 selected attributes have relation with the “Delay_Seconds” and it also makes sense as delay may be depended on factors like the date, direction, location, operator.

C) I think these attributes would also be meaningful to be considered because they indicate real life terms which contribute in causing delay: “Bearing”(direction), “VehicleRef”(a faulty vehicle). It could be possible to add other attributes like “weather” since they play role in having delay.