Part 3 - Raw data

In this part raw data processing and attribute selection is considered. Open traffic data was collected from buses in Tampere region. Each bus location is shared once a second in one batch. Data collection script polls the interface twice a second. Then data is shared through SIRI interface. In the table below collected data fields are named and explained shortly.

Attribute listing and short explanation:

'ResponseTimeDate'	Date and time of the received message from the vehicle
'ResponseTimestamp'	Seconds from first 'ResponseTimestamp' value in file
'ProducerRef'	Service producer reference
'ValidUntilTimeDate'	Date until the value may be considered valid. After this time value has passed and the bus has not sent a new data set to the background system, bus location is no longer repeated.
'ValidUntilTimeStamp'	Seconds from first 'ValidUntilTimeStamp' value in file
'RecordedAtTimeDate'	Date and time the vehicle records its location
'RecordedAtTimeStamp'	Seconds from first 'RecordedAtTimeStamp' value in file
'LineRef'	Specifies the line where the bus is currently operating
'DirectionRef'	Direction the bus is travelling on the route. 1 = from origin stop to destination stop, 2= from destination stop back to origin stop
'DataFrameRef'	Specifies the date when the vehicle started from the origin stop
'DatedVehicleJourneyRef'	Time when the vehicle started from the origin stop based on the timetable
'OperatorRef'	Operator, in Tampere region in the time of data collection: TKL, Länsilinjat (LL), Onni bussi(OB) and Paunu
'OriginName'	Name of the stop where the vehicle started
'OriginNameLanguage'	Language used to define 'OriginName'
'DestinationName'	Name of the stop where the bus is heading to
'DestinationNameLanguage'	Language used to define 'DestinationName'
'Monitored'	True/False
'VehicleLocationLongitude'	Longitude of the map position of the vehicle
'VehicleLocationLatitude'	Latitude of the map position of the vehicle
'Bearing'	Direction the vehicle is going. Angle calculated from north, 0 indicates that the vehicle is standing still.
'Delay_String'	Relative time the bus is behind or ahead of the schedule
'Delay_Seconds'	Delay string transformed to seconds (negative values for vehicle ahead of schedule)
'VehicleRef'	ID number of the monitored vehicle

What to report? Answer the following questions and report the answers as well as the required outputs in your exercise report.

Questions

1. Data cleaning

Most common cleanup procedures are: removal of incorrect information, handling missing values, noise removal, handling inconsistencies, writing to numeric, and detecting outliers. After clean up, data is integrated, transformed and reduced. The aim of the procedures is to create attributes that are easier to analyze with software, remove redundant information, and fit information in smaller space.

(a) Consider the bus data collection. How many lines of data are collected during one day if each bus location is reported once a second? Give a rough estimate.

(b) Load file rawData.arff to Weka and use *'Explorer'* to analyze the file. In theory, samples are taken twice a second, but what is the real sampling rate? There are 50000 lines of data in the file, how long did it take to collect the data? How many vehicles are included in the file (each 'VehicleRef' indicates one bus)?

(c) What are maximum, minimum and stDev of delay in seconds? Which bus line has the maximum value?

(d) What is an outlier? How does one define outliers on delay values? Where would you put a threshold and why?

(e) Explain what could be done to attributes 'OperatorRef', 'DestinationName', and 'OriginName' to make data more useable? (Use *'Edit'* and 'Filter' to make these changes to attributes)

(d) In this exercise the end goal is to consider delay analysis. For automatic analysis tasks and classification Weka needs all of the attributes to be of types: numeric, nominal, or date. Remove unusable attributes. How many attributes are left?

2. Attribute selection

(a) Use 'Select attributes' to evaluate the significance of attributes. Choose 'CfSubsetEval' as the attribute evaluator and 'BestFirst' with default settings as the search method. Select 'Delay_Seconds' as the attribute to be analyzed. How does the evaluator select the subset? How does the method evaluate significance?

(b) Run the attribute selection and report the selection. What causes them to be selected? Does the selection make sense?

(c) Consider the sample size, does it show in the selected attributes? Consider the previous answers and the selection of attributes by Weka. Which attributes would you use for delay analysis?

Last modified: 3.4.2017 by Risto Vehmas