Politecnico di Milano
Data Mining and Text Mining

Driving Style Clustering

http://code.google.com/p/dmdsc/

Filippo Sironi	734456
Matteo Villa	735064

2009

1 Data Exploration and Preprocessing

The first step of this project consisted in a preliminary exploration of the dataset to better understand its characteristics.

The idea at the very base of this process is to take advantage of human abilities for selecting an appropriate preprocessing chain and looking for dirty data that should be cleaned or discarded. Moreover, we tried to take in consideration only useful attributes that best suit the assigned task.

To ease our job throughout all the process described in this report we decided to take advantage of the RapidMiner tool.

The initial dataset contains:

- `Probe`, which is the vehicle identifier;
- `Cycle`, which identifies a period in which the engine vehicle is on;
- `Transmission` and `RawData`, which are identifiers for the GSM connection and packets;
- `Probe Date/Time`, which is the server time;
- `GPS Date/Time`, which is the car time;
- Vehicle `Speed` and `Acceleration`;
- `GPS Sat. Count`, which representes the number of satellites visibile to the GPS;
- `GPS-0`, `GPS-1`, and `Altitude`, which are respectively Longitude, Latitude, and Altitude measured by the GPS system;
- `GPS Speed`, which is the speed at which the vehicle is running according to the GPS;
- `Eng. Speed`, `Eng. Load`, `Torque`, `Fuel Rate`, `CO2 Rate`, `Eng. Temp`, which are engine parameters.

We decided to take in consideration a subset of these attributes:

- `GPS Date/Time`, renamed to `GPSDate`, is used to simplify the time series;
- `GPS-0, GPS-1`, which are used later in the mining process to find correlations between vehicle speed and geographical position;
- `Speed`, which is smoothed with a simple moving average, then averaged with respect to the window applied to the time series and finally normalized;

- `Eng. Speed`, renamed in `EngineSpeed`, which undergoes the same procedure described for the Speed attributes.

We decided not to take into account attributes like Acceleration, Torque, and Fuel Rate since they are - for real - strictly correlated to Speed and EngineSpeed and thanks to some mining test we notice that they don't have that much weight in the clustering operation.

Going deeper with the preprocessing phase we needed a work chain that is able to eliminate the noise due to the sampling of real data and all the attributes we didn't care about. In Figure 1 is shown the very first tool-chain used to in the preprocessing phase.

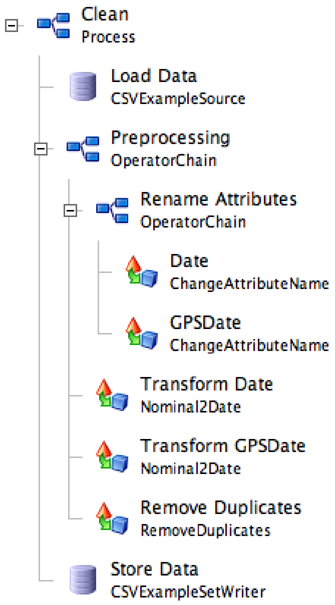


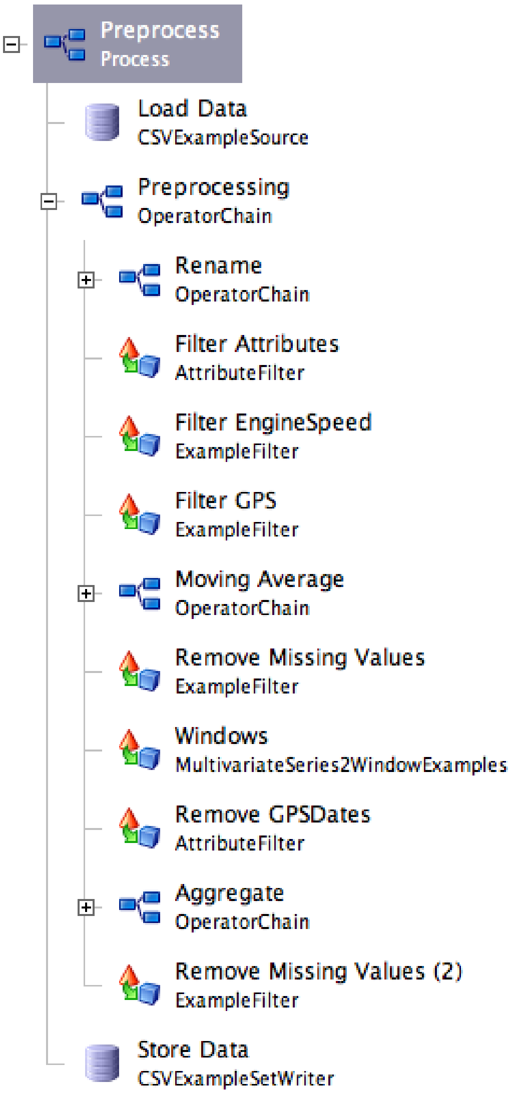
Figure 1: Clean Process

First of all, we loaded the dataset thanks to an ad-hoc CSV[1] reader operator, then we renamed date attributes and converted them in suitable format thanks to appropriate components. After these operations we removed duplicates based on the GPSDate attribute. This first phase ends with the component which is in charge of writing the resulting CSV dataset. These steps were necessary and repeated for every input dataset.

After this cleaning process we needed to work on attributes in order to select only the meaningful ones and to elaborate some of them to have suitable values for clustering. This second tool-chain is represented in Figure 2.

With this second work-flow we loaded a cleaned dataset on which we applied some attributes renaming and filter operators to select only a subset of attributes. The real preprocessing step consisted in applying the moving average on the Speed and EngineSpeed attributes in order to smooth their values. Since the moving average works on a sliding set of values and some of these are missing, we needed to prune the corresponding instances from the dataset before proceeding with the operations.

[1] Comma Separated Values



- Preprocess
  Process
  - Load Data
    CSVExampleSource
  - Preprocessing
    OperatorChain
    - Rename
      OperatorChain
    - Filter Attributes
      AttributeFilter
    - Filter EngineSpeed
      ExampleFilter
    - Filter GPS
      ExampleFilter
    - Moving Average
      OperatorChain
    - Remove Missing Values
      ExampleFilter
    - Windows
      MultivariateSeries2WindowExamples
    - Remove GPSDates
      AttributeFilter
    - Aggregate
      OperatorChain
    - Remove Missing Values (2)
      ExampleFilter
  - Store Data
    CSVExampleSetWriter

Figure 2: Preprocess

2 Mining Process

2.1 Clustering

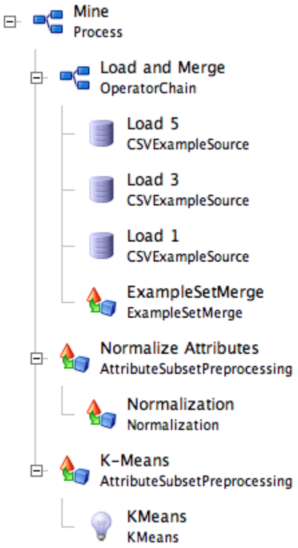
Since all the data coming from the preprocessing are unlabeled, we chose an unsupervised learning approach for our mining process (shown in Figure 3); so we start with a clustering operation on all the data.

After the preprocessing step we have a distinct file for each car involved in the test, so we have to merge all of them.

Once all the data are stored in a single location, we can normalize the numerical attributes Speed and EngineSpeed, in order to make them weight the same during the clustering process.

Finally the clusters are computed, using the k-means algorithm with 3 centroids; this number of final clusters seemed to provide the better result.

Figure 4 shows the cluster distribution with respect to the Speed and EngineSpeed attributes.



Mine
Process
Load and Merge
OperatorChain
Load 5
CSVExampleSource
Load 3
CSVExampleSource
Load 1
CSVExampleSource
ExampleSetMerge
ExampleSetMerge
Normalize Attributes
AttributeSubsetPreprocessing
Normalization
Normalization
K-Means
AttributeSubsetPreprocessing
KMeans
KMeans

Figure 3: Mine Process

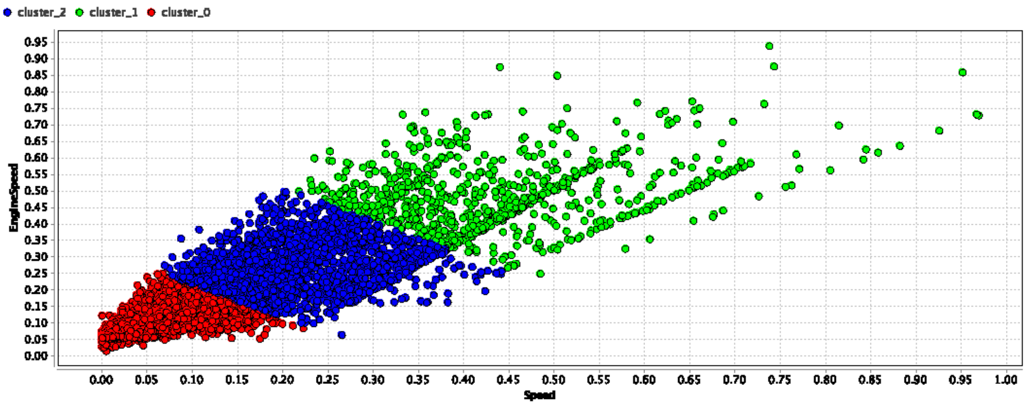


Figure 4: Clustering result