Politecnico di Milano
Data Mining and Text Mining

Driving Style Clustering

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

Filippo Sironi 734456
Matteo Villa 735064

2009

Contents

List of Figures

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 discorded. Moreover, we tried to take in consideration only useful attributes that best suit the assigned task.

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;
- `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;
- `Speed`, which is smoothed with a simple moving average, normalized, and then averaged with respect to the window applied to the time series;
- `Eng. Speed`, renamed in `EngineSpeed`, endures 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 preprocessing 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.

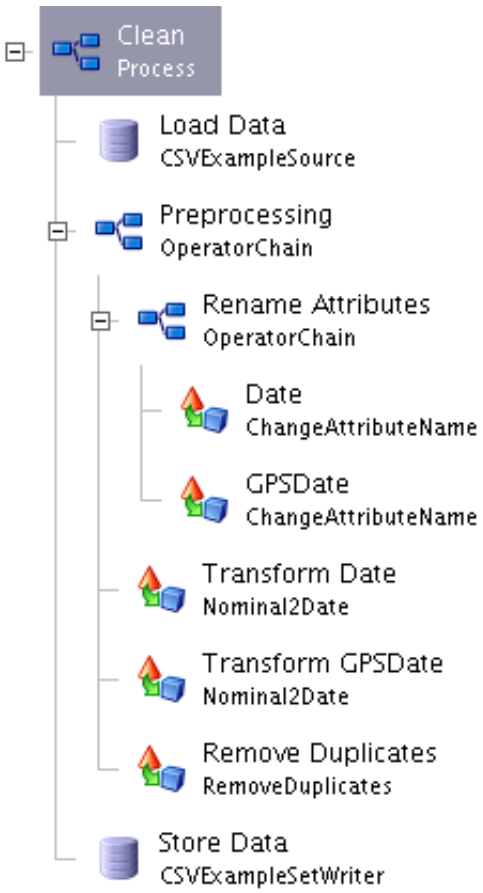

Figure 1: Clean Process

First of all, we loaded the dataset thanks to an ad-hoc CSV [1] component, 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. The tool-chain ends with the component which is in

[1] Comma Separated Values

charge to write the 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 **??**.

With this tool-chain we loaded a cleaned dataset on which we applied some renaming and filter operators before starting the real preprocessing. The preprocessing 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 some of computed values are missing so we needed to prune them from the dataset. Finally, our preprocessing contained also a normalization operator to make every value weight the same during the clustering process.