

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 represents the number of satellites visible 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 GPSTime, 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.

First of all, we loaded the dataset thanks to an ad-hoc CSV¹ 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 GPSSDate 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 then first results are missing, more precisely, the number of missing values is the window dimension minus one. Since we didn't want to deal with these missing values we pruned

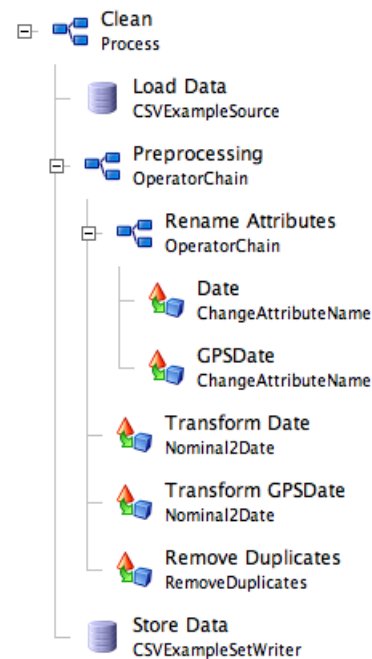


Figure 1: Clean Process

¹Comma Separated Values

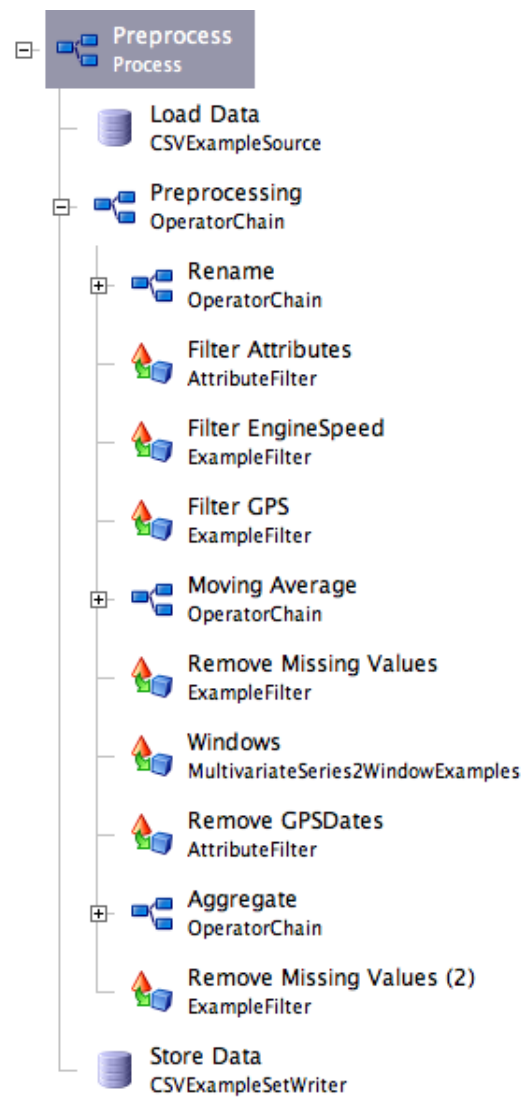


Figure 2: Preprocess

the corresponding examples from the dataset before proceeding with the next operations.

2 Data Mining

Since all the data coming from the preprocessing are unlabeled, we chose an unsupervised learning approach for our mining process which is shown in Figure 3; hence we started with a clustering operation on all the data.

After the preprocessing step we had a distinct file for each car involved in the test, so we had 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 distance calculation needed by the clustering algorithm.

Finally the clusters are computed using the K-Means algorithm initialized with 3 centroids: we tested the clustering algorithm with different number of centroids and we chose this number of clusters since it seemed to provide reasonable results in terms of partitioning.

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

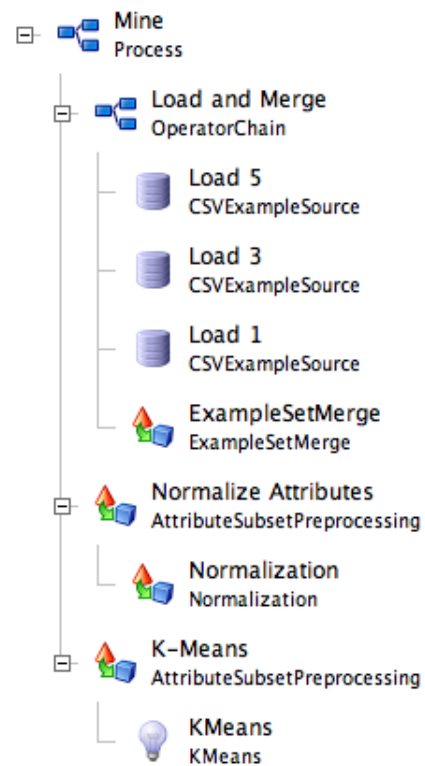


Figure 3: Mine Process

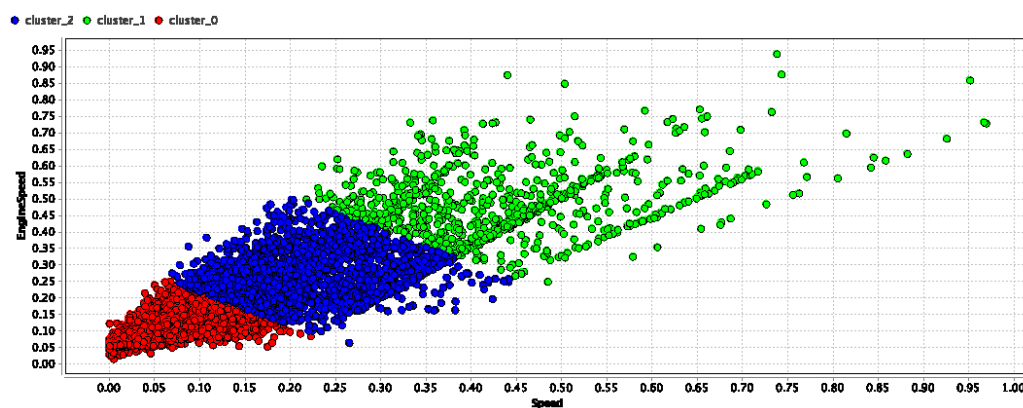


Figure 4: Clustering result

From Figure 4 we can easily deduce that Speed and EngineSpeed are cor-

related by means of a “linear” function. In fact Figure 5 and Figure 6 which represents the distribution of Probes over clusters with respect to Speed and to EngineSpeed are very similar.

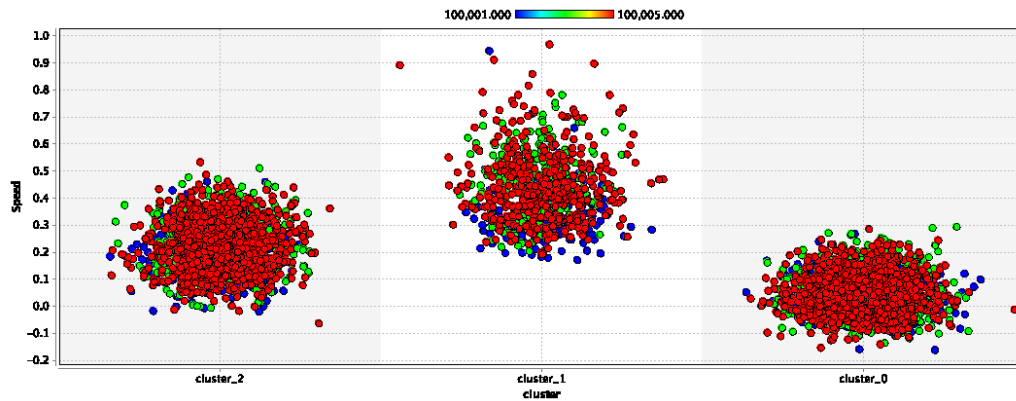


Figure 5: Probes distribution over clusters with respect to Speed

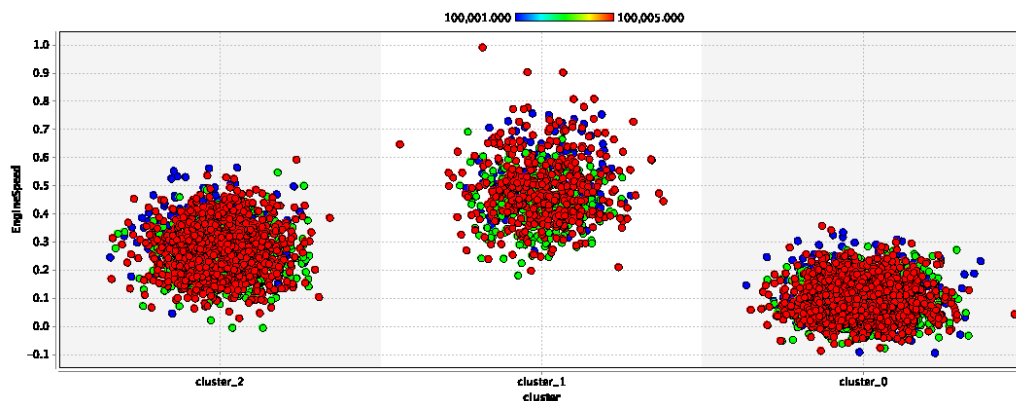


Figure 6: Probes distribution over clusters with respect to EngineSpeed

From both Figure 5 and Figure 6 we can deduce that all vehicles assume all the three driving styles the mining process discovered.

Figure 7 shows an interesting trend: clusters, which basically describe how a vehicle behaves, seems not to be correlated with GPS coordinates hence with the portion of the city in which the vehicle moves since all clusters are distributed over the cartesian graph.

More information on this trend can be extracted from Figure 8, 9, and 10 which are related to the three vehicles we analyzed.

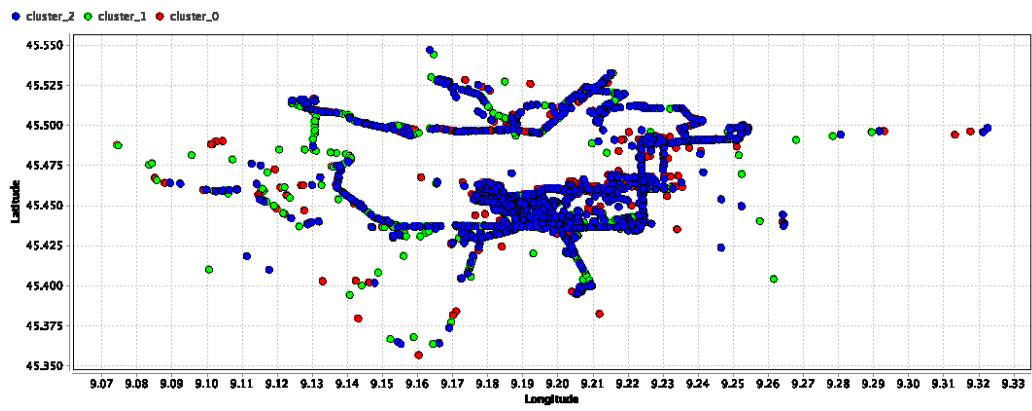


Figure 7: Cluster distribution over GPS coordinates

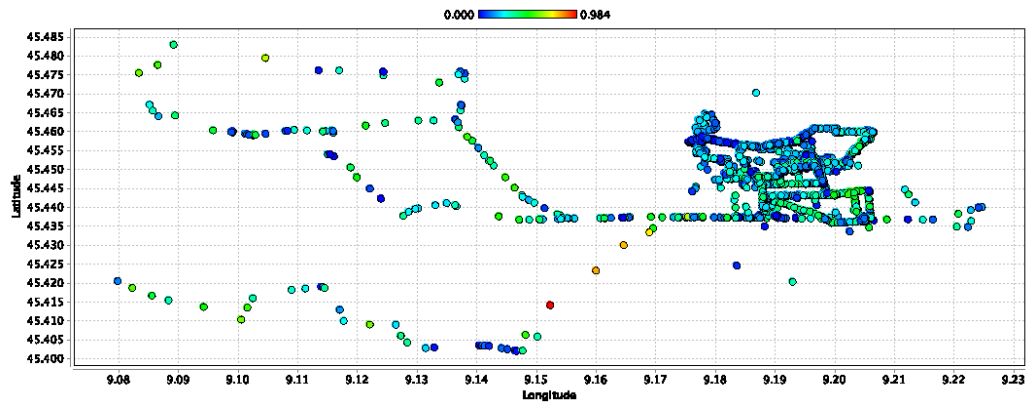


Figure 8: Speed distribution over GPS coordinates

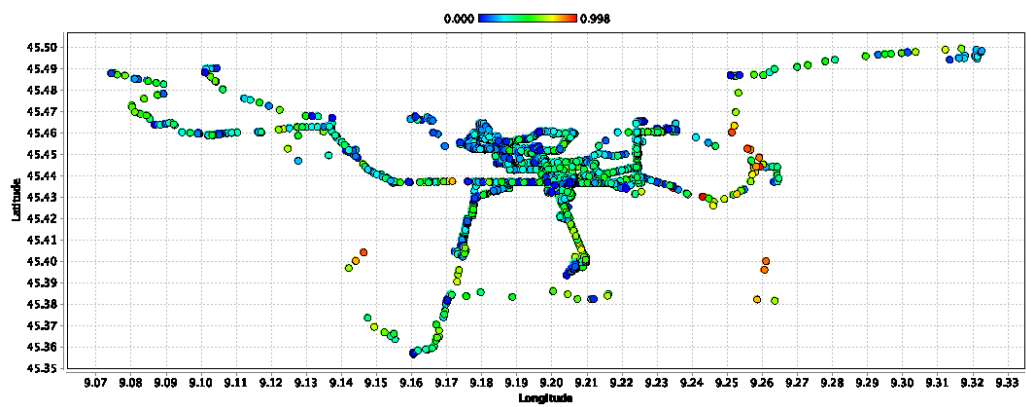


Figure 9: Speed distribution over GPS coordinates

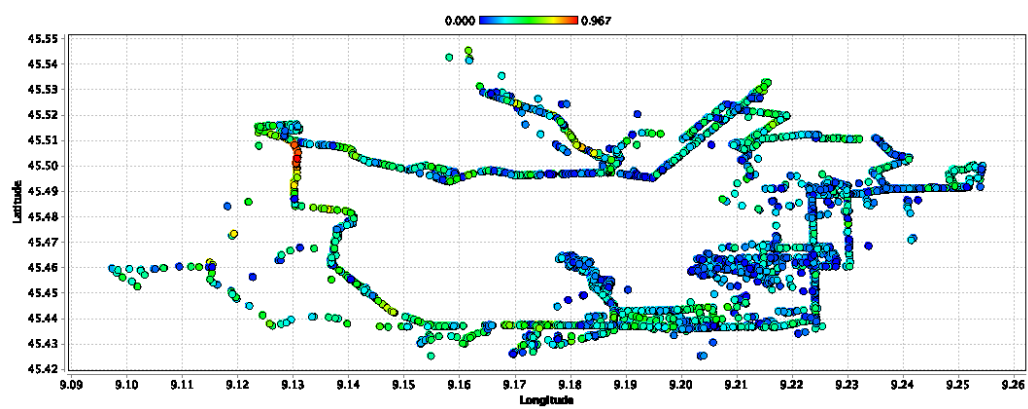


Figure 10: Speed distribution over GPS coordinates