Driving Behavior Analysis: An Approach Using Clustering Algorithms

Lucas P. Chabariberi [1], **Francisco N. C. Sobral**[2], **Sarajane M. Peres**[1]

[1]Escola de Artes Ciências e Humanidades – Universidade de São Paulo (USP)
São Paulo – SP – Brazil

[2]Departamento de Matemática – Universidade Estadual de Maringá
Maringá – PR – Brazil

{lucaschabariberi,sarajane}@usp.br, fncsobral@uem.br

Abstract. *The discovery and characterization of driving behavior profiles can be useful to support the optimization of processes for insurance companies or fleet managers. The evolution of ubiquitous computing and data analysis techniques have made such tasks possible. In this paper, we present a study on the analysis of driving behavior through clustering algorithms on a real-world dataset. We applied the k-Means++ and Spectral Clustering algorithms that came up with results that showed: the existence of less and more aggressive behavior profiles; potential to discover more profiles since preliminary quantitative evaluation indicated good quality in clustering with four profiles.*

1. Introduction

The development of new technologies made it easy to collect data from drivers and vehicles. Car insurance and fleet managing companies have increasingly interest in analyzing this data. They use smartphones or specific devices inside vehicles to obtain information about trips. Insurance companies offer discounts to "good" drivers, while fleet managers identify driving patterns that could result in accidents or damage to vehicles. In both cases, identifying risky driving behaviors can lead to several economic benefits.

Unbiased detection of safe or risky driving behavior is a difficult task. For fleet managers, the task is even harder, since a car can have several drivers. Therefore, the use of unsupervised learning is desirable. In [Constantinescu et al. 2010], clustering techniques have been applied to identify groups of drivers, using information obtained by a in-house device. The data was analyzed by Principal Component Analysis (PCA) and by hierarchical clustering. In [Castignani et al. 2015] information obtained by a smartphone was used to develop a driver-score system. The driver starts with score 100 and each event reduces this score by a predefined amount. PCA and clustering techniques were used to assess that the score system was able to correctly identify different driving behaviors.

In this paper, we analyze data obtained from devices connected to vehicles by the company Cobli[1]. The goal is to identify groups of driving behaviors, instead of groups of drivers. Six months of data, collected from one device was analyzed using the k-Means++ and Spectral Clustering algorithms and further explored with the Silhouette index and features-by-clusters distribution graphics. In Section 2 we provide the background of these algorithms and the Silhouette index. In Section 3, the research context is detailed. In Section 4 we discuss the experiments and the results. Conclusions are drawn in Section 5.

[1]Cobli is a company that develop telemetry and management system for fleets: https://cobli.co/

2. Background

k-Means++ algorithm: k-Means creates a partition for the dataset in which k clusters of data are established. It implements an optimization process, in which intraclusters similarity is minimized while the interclusters similarity is maximized, trhough the steps: k initial centroid vectors are randomly generated to represent k clusters; each datapoint is associated with the centroid most similar to it; each centroid is updated to become the average vector of the datapoints associated with it. The parameter k and the distance metric are usually determined by a data analyst [Silva et al. 2016]. Formally, let a dataset $X = \{\overrightarrow{x_i}\}$, with $i = \{1, ..., n\}$ to be partitioned in k clusters $G = \{g_j\}, j = \{1, ..., k\}$. k-Means searches for a partition in X that minimizes $P(U, G) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ji} dist(\overrightarrow{g_j}, \overrightarrow{x_i})$, in which $\overrightarrow{g_j}$ is the centroid for g_j, $U = \{0, 1\}^{k,n}$ is an indicator matrix that associates each $\overrightarrow{x_i}$ to one $\overrightarrow{g_j}$, and $dist$ is a distance metric. k-Means++ is an extension in which a datapoint $\overrightarrow{x} \in X$ is randomly chosen as the first centroid and the next ones are chosen following the probability $dist_{min}(\overrightarrow{x_l}) \ / \sum_{\overrightarrow{x} \in X} dist_{min}(\overrightarrow{x})$, with $dist_{min}(\overrightarrow{x})$ as the distance between datapoint $\overrightarrow{x} \in X$ and the nearest centroid already chosen [Arthur and Vassilvitskii 2007].

Spectral clustering: The spectral clustering is a modern clustering technique. Given X the dataset and k the number of desired clusters, the first step is to build the *similarity matrix* $S = (s_{ij})$, where s_{ij} represents the similarity between datapoints $\overrightarrow{x_i}$ and $\overrightarrow{x_j}$, $i, j = 1, \ldots, n$. Using S, the *similarity graph* and its associated *graph Laplacian* L are constructed. The graph Laplacian is formed by matrix operations using the *weighted matrix* $W = (w_{ij})$ and the *degree matrix* D. W is constructed using information from the similarity graph, where w_{ij} is zero, if there is no edge connecting datapoints $\overrightarrow{x_i}$ and $\overrightarrow{x_j}$ (represented as vertices of the graph), or a positive value given by s_{ij}, if there is an edge between the vertices i and j. Matrix D is a diagonal matrix such that $d_i = \sum_{j=1}^{n} w_{ij}$ represents the (weighted) degree of vertex i. Different calculations involving W and D result in graph Laplacians with different properties. The k eigenvectors $\{\overrightarrow{u_1}, \ldots, \overrightarrow{u_k}\}$, associated with the k smallest eigenvalues of L, are then calculated and used to build matrix U. The last step is to consider the *rows* of U as the new datapoints and apply k-Means to find k clusters. Several different strategies can be used at each step of this technique, from the way that similarity is calculated to the clustering algorithm in the transformed dataset. Theoretical details and the different strategies are discussed in [von Luxburg 2007].

Silhouette index: The Silhouette index I_S is an internal validation index used to evaluate clustering quality [Rousseeuw 1987]. It implements an internal validation because only the distribution of datapoints is considered in the index calculation. This type of index evaluates the relation between the compactness and the separability of the clusters. First, the I_S index is calculated for each datapoint; then, it is calculated for each cluster as the average of I_S related to the datapoints in the same cluster; or for the whole dataset as the average of I_S related to all datapoints. It index is limited to $[-1, 1]$. The closer to 1, the better defined the clusters are, and the closer to -1, the opposite is observed, leading to the conclusion that the organization of datapoints in the clusters is wrong [Rousseeuw 1987].

3. Research context

Data collection was done through a device called OBD ("On Board Diagnostics") coupled to each vehicle of the fleet. It collects various diagnostic information and sends it to the

office server as events. The frequency of sending varies from seconds to minutes, depending on the status of the vehicle (e.g., turned on or off). Four pieces of information were used: *deviceId*, the OBD identification; *timestamp*, the event's date and time; *ignition*, the vehicle status; and *eventCode*, the event type. Events types are divided into two groups: risk and non-risk. A risk event indicates an aggressive driving behavior, for example, driving at dangerous speed, performing quick accelerations, sharp turns or breaks. Non-risk events are related with sensor control information, vehicle status, vehicle location, among others. Herein, we analyze driving behavior profiles by using the event type field.

4. Experiments and results

Dataset and preprocessing: Table 1 describes three dataset versions. The dataset was collected from January to July, 2018. The original dataset is bigger than that used in our experiments. A reduction was performed to allow an efficient study on the feasibility of clustering approaches in driving behavior analysis: only data from one vehicle was used.

Table 1. Dataset versions description

Dataset	risk events	non-risk events	total of events	vehicles	drivers	trips
Original	63.355	3.478.088	$\approx$ 3.5 millions	20	38	13.319
Reduced	2.780	129.722	132.502	1	3	461
Preprocessed	2.758	118.110	120.868	1	3	461

We executed three preprocessing steps on the reduced dataset: selection of events of interest; data aggregation by using timestamps; and normalization. In the first step, we considered all events for one *deviceId* that occurred during an actual trip (events received while the vehicle was off were discarded). Then, the datapoints were organized as vectors (rows of the data matrix) of attributes (columns of the data matrix). Each attribute of a datapoint corresponds to the number of times that each type of event occurred in one day[2]. Finally, we applied a linear transformation on each attribute in order to map them to $[0, 1]$ and removed the column associated with non-risk events. The final dataset comprises 68 datapoints represented by 11-dimensional vectors, organized in a 68×11 data matrix.

Setup of experiments: Both algorithms were carried out with $k = \{2, 3, 4\}$, Euclidean distance and stop conditions based on maximum number of iterations or small quantization error, which happens first. The Spectral clustering was carried out considering the Gaussian kernel with $\sigma = 1$ for similarity and k-Means for clustering phase.

Results and analysis: The clustering results were evaluated considering three aspects: index silhouette values, index silhouette graphs and features-by-clusters distribution graphics. Figure 1 presents a visual summary of the results obtained with each algorithm. For each algorithm, the figure shows the Silhouette values and graphics for each value of k and examples of features-by-clusters distributions for $k = 2$ and three features.

In terms of Silhouette index and comparing both algorithms, we observed slightly better results with k-Means++, for $k > 2$. In k-Means++ results, variations in k values do not produced significant differences. However, for $k = 4$, a smaller amount of datapoints achieved negative values for I_S. Spectral algorithm presented a greater difficulty in

[2]According to the data owners, it is fair to assume that a vehicle is driven by one driver during one day.

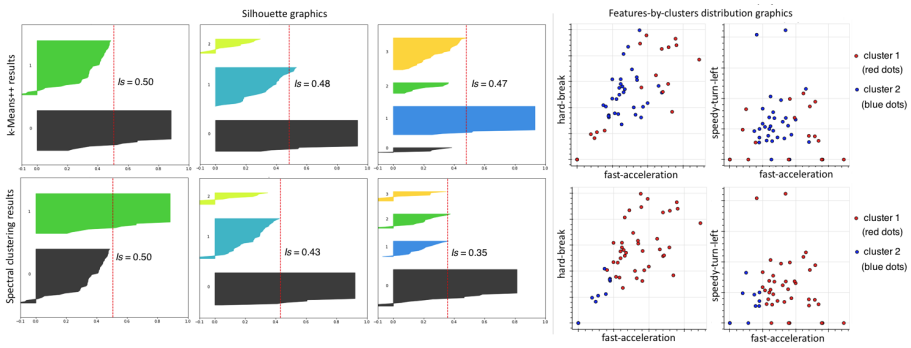

Figure 1. Visual summary of the results

organizing the clusters, as can be proved through the number of datapoints that have negative values for I_S. The results were also analyzed in terms of the distribution of features by clusters to provide a semantic interpretation. The clusters built by k-Means++ were hard to interpret due to the fuzzy boundaries established between clusters. For Spectral clustering, similar situations were observed for $k > 2$. However, Spectral clustering with $k = 2$ allowed to characterize clusters based on three features: fast-acceleration, hard-break and speedy-turn-left. We claim that these three features are highlighting the main characteristics of aggressive driving behavior in the context of our sample dataset.

5. Conclusions

The results presented herein showed the suitability of clustering for discovering and explaining driving behaviors. The next steps comprise more detailed analysis on the obtained results, and the extension of the experiments for the complete database.

Acknowledgment

The authors thank *Cobli - Sistema de Telemetria e Gestão de Frotas* - for the partnership and financial support in the development of this research.

References

Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Proc. of the 18th annual ACM-SIAM Symp. on Discrete Algorithms*, pages 1027–1035.

Castignani, G., Derrmann, T., Frank, R., and Engel, T. (2015). Driver Behavior Profiling Using Smartphones: A Low-Cost Platform for Driver Monitoring. *IEEE Intelligent Transportation Systems Magazine*, 7(1):91–102.

Constantinescu, Z., Marinoiu, C., and Vladoiu, M. (2010). Driving Style Analysis Using Data Mining Techniques. *Int. J. of Computers Communic. & Control*, 5(5):654–663.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. of Comp. and Applied Math.*, 20:53–65.

Silva, L. A., Peres, S. M., and Boscarioli, C. (2016). *Introdução à mineração de dados: com Aplicações em R*. Elsevier.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.