Classifying Vehicles' Behaviors using Global Positioning Systems Information

Alessandro Silacci*, Julien Tscherrig*, Elena Mugellini* and Omar Abou Khaled*
Emails: alessandro.silacci@hes-so.ch, julien.tscherrig@hes-so.ch,
elena.mugellini@hes-so.ch and omar.aboukhaled@hes-so.ch
*HES-SO, University of Applied Sciences and Arts Western Switzerland

***Abstract*—This study presents a solution to enhance the cities' traffic control by classifying particular vehicles' behaviors. A Support Vector Machine (SVM) approach is presented, enabling the system to classify cars that are looking to park and those that are simply transiting through a city. Through this paper, we also propose a new way of managing the high density of traffic data using a grid. The results show that the system is able to distinguish the two different behaviors with an accuracy averaging 80%.**

Keywords–Machine learning; Intelligent Traffic System; Feature Selection; Global Positioning System

I. INTRODUCTION

Today's big cities are getting submerged by the traffic growth. It started to become critical as the traffic grew and the cities are now facing problems like traffic congestion and others, although a desperate try to mitigate them. Thus, Intelligent Transportation Systems (ITS) and research in that sense, are also getting more attention. Most cities struggle to counteract such issues due to the lack of flexibility in their architecture. The use of ITS solutions is therefore helpful when a city tries to understand the possible bottlenecks or other particular behaviors observable in its streets. Through these identification capabilities the cities are trying to solve the problems linked to their architecture.

We based our research on [1] data, which provides the Global Positioning System (GPS) location of many smartphones at a given interval in the city of Aracaju (Brazil). The dataset is composed of either people simply walking, taking a bus or a car. Using such location technology has been motivated by the wide amount of device capable of using it, considering the growth of the smartphone market.

The following study is oriented towards enriching a city's knowledge of its traffic by differentiating the parking patterns of the transiting patterns by extracting it from the Global Positioning System's data of a moving vehicle. The Section III show details about the dataset and the features that we selected or created, our system is explained in the Section III-C, its results and a final discussion are presented in Section IV and Section V respectively.

II. RELATED WORK

Behavior detection or classification is a field where many research projects are trying to respond the best they can, it is even defined as the field having the least research by [2]. In the agricultural domain, farmers are trying to understand how their cattle behave during a certain period of time. Therefore, they use specific sensors but also couple them with the GPS information they get [3]- [5].

Researchers identified that the raw GPS points cannot be used very efficiently. Therefore, information like the speed, direction or distance was inferred from the GPS records and the time of the capture [6]. Most of the reviewed papers would rather focus their models on determining the users' actual activity across the time. This means that they were first trying to determine if the participant was moving, if so what the transportation mode he was using. [7] proposed a system based on fuzzy logic to identify if a person was walking, taking some sort of transportation method or simply staying. Their model is capable of determining the three situations according to the angle between the points and the speed computed from it. Based on few features, the model is capable of giving the probability of the actual segment to pertain to one of the three sets.

A machine learning approach was used in an article from [8] which labelled multiple segments of a GPS trace. Labelled segments were checked for errors using some fuzzy logic rules, ending in a multi-staged technique to provide the corresponding label to a segment. The usage of SVM was helpful in classifying the mode of transportation used, since the model was capable of identifying four types of vehicle types (car, bus, train or tram). [9] also proposed a multilayered classification model, composed of a decision tree and a Hidden Markov Model (HMM). Using their pipeline, they were able to categorise transportation modes like running, walking, biking, stationary position and motorised transport (no distinction).

The work from researchers in [10] are exposing different ways of training an Artificial Neural Network for pattern recognition, and the results tend to demonstrate that this depends on the amount of data available. [11] developed a neural network, based on a multi-layer perceptron model that is capable of identifying the mode of transportation used by a person. They offer, through a mobile application called *TRACIT*, the capabilities to determine when a trip began and finished and how the person travelled in order to enhance surveys' processes and ease their use. They also note the importance of certain features in determining the type of transportation, like the acceleration and the total distance of the trip. The previously presented systems were beaten by *TrajectoryNet*, a Recurrent Neural Network (RNN) model proposed by [12]. It was able to classify GPS trajectories in four categories; bike, car, walk and bus, with the precision of 98% which is beating at least by 4% the results of [8]. Similarly, [13] presented two different approaches of managing the raw GPS tracking points. They used fuzzy logic rules and a random forest model in order to recognise indoor and in-vehicle travels and mentioned some problems to determine if a pedestrian was walking or not.

Researchers like [14] also tried to predict potential accidents or traffic congestion based on data inferred from the GPS location of vehicles. They segmented roads using various points along them and then computed different variables relatively to them. [15] were able to extract and classify three traffic congestion levels using a Decision Tree algorithm with

 ISBN: 978-1-61208-685-9

a high accuracy. In order to assess highway traffic conditions, [16] have demonstrated that the SVM and information inferred from vehicles' positions could provide results above 75%.

III. METHODS

A. Data Generation

UCI's dataset is composed of GPS points series that describe the movements of people using their mobile phone GPS antenna. This statement induces that the data had to be filtered while being labelled. The GPS series were displayed and labelled visually, meaning that we relied on the actual position of paths' segments in order to attribute a specific label to them. Going through such process also helped to identify paths that were done by people walking and not using a vehicle.

The second step in the process was to create an appropriate tracks' dataset. Therefore, we decided to create a grid around the city center. Each of the cell is covering a parcel of a hundred meters squared and the total coverage of the city was ten thousand meters squared. We then computed for each grid cell the number of unique tracks going through it. This count helped us understand and filter the useless parts of the grid, reducing the total matrix. Tracks were then processed to produce the dataset which contains four different features. We decided to keep the track_id (which was unique for each record), the "from_win_id" which is the origin's cell id of a track, the "to_win_id" which is the destination cell id, and finally "delta_t" the time delta between the origin to the destination in seconds. An example of the grid with a path is given in the Figure 1 and its matching Table I.

All the records were labelled using two different classes: "transit" meaning the vehicle is transiting through the city and "parking" identifying a car looking for a parking. A path could be segmented with both transiting and parking segments. We additionally created a dataset filtering the paths containing at least two different classes ("parking" or "transit").

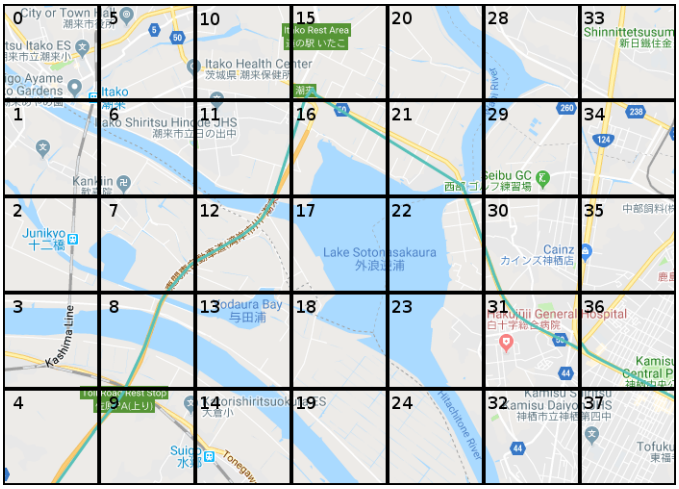


Figure 1. A path (turquoise) going through the grid.

B. Data Analysis

The dataset generated includes a few features, but to assess their variance and to ensure a good usage of them during the training and testing phase we decided to run multiple feature selection techniques. We first tried to understand the variance ratio of each feature and so understand their importance in the future model usage. We therefore used a Principal Component Analysis (PCA) to get a clear observation of the ratios, as demonstrated in Figure 2. Obviously, the track_id revealed to be the most important feature, but we will not use it as it may skew the results by relying too much on this value which may be dynamic in future usage. Another observation from Figure 2 is that the delta between the cell movement has a lot of same values, but this is understandable since the vehicles could take the same time to transit through cells. We further explored with two types of feature selection techniques, the filters and the wrappers. Filters are based upon statistical measure results and provide a score for each feature, while wrappers are using machine learning models in order to determine a feature ranking depending on the score obtained by the classifier using a particular set of them.

TABLE I. SAMPLE OF THE GENERATED DATASET FOR A PATH.

Track_id	*From*	*To*	*Delta_t*
1	4	9	37
1	9	8	60
1	8	12	200
1	12	11	85
1	11	15	40
1	15	21	172
1	21	22	29
1	22	30	14
1	30	31	56
1	31	36	193
1	36	37	230

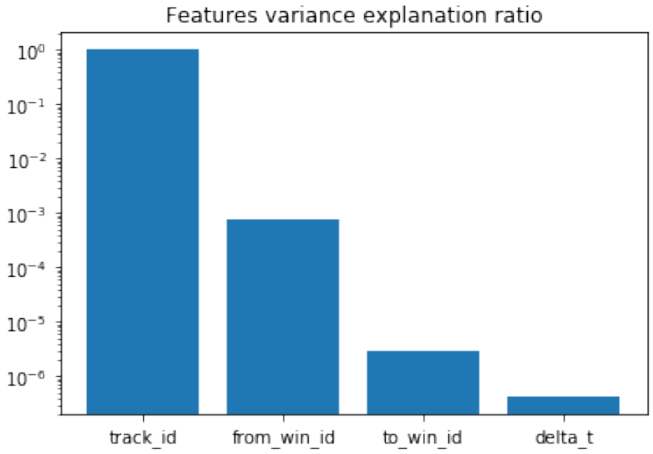


Figure 2. Principal Component Analysis of the dataset in a log scale.

We started with the filters and therefore selected the ANOVA-F measure as demonstrated in Figure 3 and the χ^2 which is observable through Figure 4. The results provided by the two measures are slightly different, especially looking at the track_id and delta_t features which are interpreted differently. This is due to their specific characteristics, where

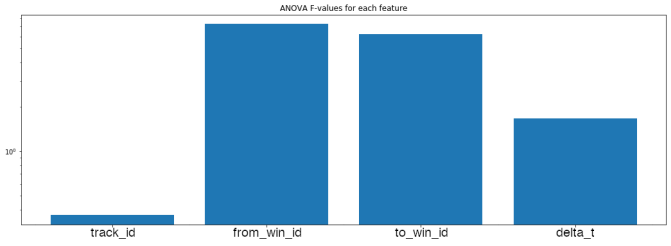


Figure 3. The ANOVA-F feature selection scores.

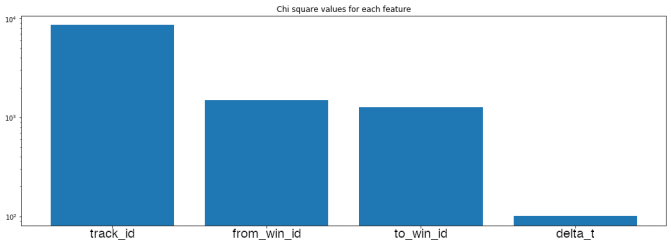


Figure 4. The χ^2 feature selection scores.

the χ^2 is a measure taking more the number of times the same observations is made between a feature value and the class and ANOVA-F is scoring features by analysing the variance of Fisher's test values.

As for the wrappers, we used a Decision Tree where we selected the two best features. The model and approach is based on the scikit-learn python library [17]. Using the model, we obtained a new histogram for the Decision Tree as shown in and Figure 5. By omitting the track_id feature from Figure

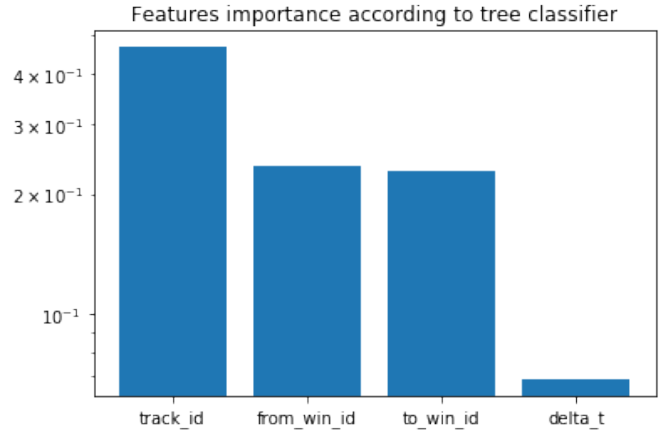


Figure 5. Decision Tree feature selection results.

2, Figure 3, Figure 4 and Figure 5, one can still make the observation that the origin cell is determinant in the dataset generated.

C. Model Selection

We ran the whole dataset in a pipeline composed of multiple classifiers. We dedicated 70% of the dataset to be the training set and used it to make a 5-fold cross-validation. Results of the classifications were then compared as in Figure 6. We observed the standard deviation of the results and the accuracy for each classifier and came to the conclusion that the SVM was the promising model to use.

We decided to further explore the SVM classifier and made the fine-tuning of its hyper-parameters using a grid searching approach. The algorithm was given four different parameters range configurations and used a 3-fold cross-validation. The best scores were obtained using the following parameters:

- Kernel: Radial Basis Function (RBF)
- C: 1.0
- Tolerance: 0.001

IV. Results

We ran each dataset through the feature selection techniques described earlier and compared the results. At each

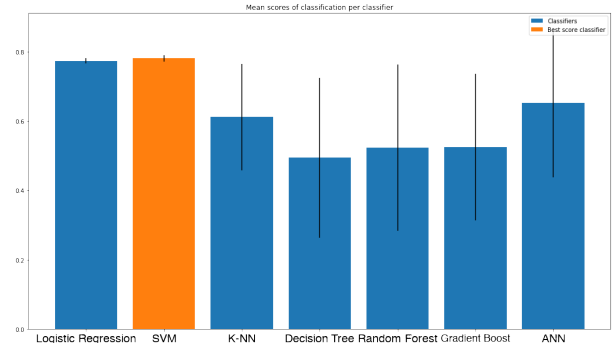


Figure 6. The accuracy results of all the tested classifiers and their standard deviation.

step, the resulting dataset containing only the selected features were used to classify the 30% test set that has never been used during the previous steps.

By comparing the two confusion matrices in Figure 7a and Figure 7b resulting from the classification using the whole set of features, we observed that the filtered paths' classification was harder. The model is providing 3.56% of false positive for the transit class while for the whole dataset it represented only 0.06%. We observed an inverted tendency on the false positives of the parking class with this error representing 16.35% and 13.91% respectively for the non-filtered and the filtered dataset. This is certainly due to an unbalance between the number of transit records and the parking ones.

By running the results of each feature selection subsets we observed that the best-performing one was the whole dataset. This is demonstrated by Table II, which summarizes different scores of three different measures. Reducing the features did not enhance the results, even if the results are still good. As the paths were entirely present in either the training or testing set, using all the features was not skewing the algorithm. The results also demonstrate that our model is capable of providing a classification precision of 86%.

TABLE II. Comparison of the results obtained by running all the different subsets of each dataset.

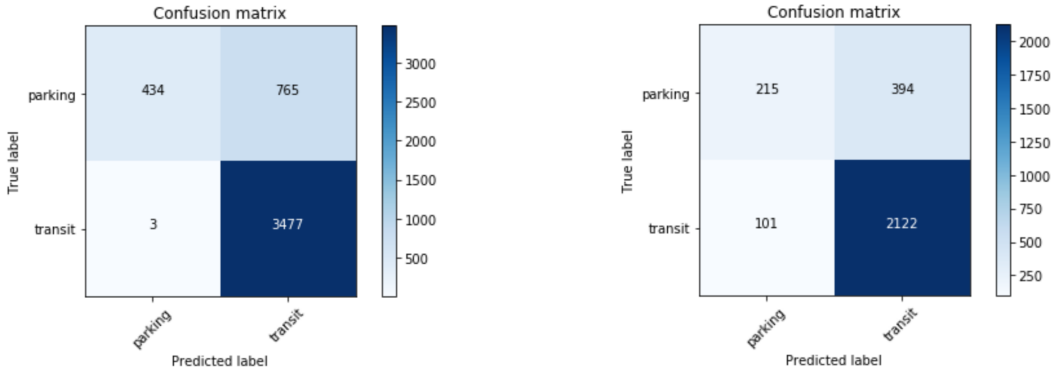
Dataset	*Precision*	*Recall*	*F1-score*
all features, all paths	86	0.84	0.81
all features, filtered paths	80	0.81	0.79
2 best tree features, all paths	83	0.81	0.78
2 best tree features, filtered paths	75	0.75	0.71
2 best ANOVA-F features, all paths	77	0.79	0.78
2 best ANOVA-F features, filtered paths	75	0.72	0.63
2 best χ^2 features, all paths	78	0.81	0.77
2 best χ^2 features, filtered paths	77	0.78	0.75

V. Conclusion

To conclude our study, we demonstrated a new approach attempting to answer the problematic linked to one of the least researched domains in Intelligent Transportation Systems. We were able to obtain significantly good results in detecting whether a car was looking for a parking place or simply transiting through the city. We compared results of many different subsets, using techniques normally used for dimensionality reduction. Unfortunately, these subsets did not provide better results, but at least we identified that a key factor in the decision-making was the origin of the vehicle, if we ignore the path's id.

Results might even be further enhanced by a better annotation technique, and a better data collection quality as the

 ISBN: 978-1-61208-685-9



(a) Using the whole dataset.

(b) Using the filtered dataset.

Figure 7. Comparison of the confusion matrices.

UCI dataset was based on people's smartphone location system and not directly from cars. This will be the subject of a future work as this study is part of an HES-SO directed project named Mobicam, specifically meant to solve the data collection and data treatment problematic.

Further work could also investigate more classes for behaviors like stopping to get somebody, waiting for somebody or traffic congestion.

ACKNOWLEDGMENT

The research program Mobicam is part of the large-scale thematic research programs of the University of Applied Sciences and Arts of Western Switzerland (HES-SO). The authors gratefully acknowledge funding from HES-SO.

REFERENCES

[1] M. O. Cruz, H. T. Macedo, R. Barreto, and A. P. Guimarães, "UCI gps trajectories data set," 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/GPS+Trajectories

[2] S. Sivaraman and M. M. Trivedi, "Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis," IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 4, Dec. 2013, pp. 1773–1795.

[3] E. D. e. a. Ungar, "Inference of Animal Activity From GPS Collar Data on Free-Ranging Cattle," Rangeland Ecology & Management, vol. 58, no. 3, May 2005, pp. 256–266. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1550742405500344

[4] S. e. a. Grünewälder, "Movement Activity Based Classification of Animal Behaviour with an Application to Data from Cheetah (Acinonyx jubatus)," PLOS ONE, vol. 7, no. 11, Nov. 2012, p. e49120. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049120

[5] M. Schwager, D. M. Anderson, Z. Butler, and D. Rus, "Robust classification of animal tracking data," Computers and Electronics in Agriculture, vol. 56, no. 1, Mar. 2007, pp. 46–59. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169907000026

[6] M. Boukhechba, A. Bouzouane, B. Bouchard, C. Gouin-Vallerand, and S. Giroux, "Online Recognition of People's Activities from Raw GPS Data: Semantic Trajectory Data Analysis," in Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments, ser. PETRA '15. New York, NY, USA: ACM, 2015, pp. 40:1–40:8. [Online]. Available: http://doi.acm.org/10.1145/2769493.2769498

[7] N. Wan and G. Lin, "Classifying Human Activity Patterns from Smartphone Collected GPS data: A Fuzzy Classification and Aggregation Approach," Transactions in GIS, vol. 20, no. 6, Dec. 2016, pp. 869–886. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12181

[8] L. Zhang, S. Dalyot, D. Eggert, and M. Sester, "Multi-stage approach to travel-mode segmentation and classification of gps traces," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: [Geospatial Data Infrastructure: From Data Acquisition And Updating To Smarter Services] 38-4 (2011), Nr. W25, vol. 38-4, no. W25, 2011, pp. 87–93. [Online]. Available: https://www.repo.uni-hannover.de:443/handle/123456789/1167

[9] S. e. a. Reddy, "Using Mobile Phones to Determine Transportation Modes," ACM Trans. Sen. Netw., vol. 6, no. 2, Mar. 2010, pp. 13:1–13:27. [Online]. Available: http://doi.acm.org/10.1145/1689239.1689243

[10] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks," Pattern Recognition, vol. 40, no. 1, Jan. 2007, pp. 4–18. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320306002081

[11] P. e. a. Gonzalez, "Automating Mode Detection Using Neural Networks And Assisted GPS Data Collected Using GPS-Enabled Mobile Phones," New York, p. 12.

[12] X. e. a. Jiang, "TrajectoryNet: An Embedded GPS Trajectory Representation for Point-based Classification Using Recurrent Neural Networks," arXiv:1705.02636 [cs], May 2017, arXiv: 1705.02636. [Online]. Available: http://arxiv.org/abs/1705.02636

[13] J. Wu, C. Jiang, D. Houston, D. Baker, and R. Delfino, "Automated time activity classification based on global positioning system (GPS) tracking data," Environmental Health, vol. 10, no. 1, Nov. 2011, p. 101. [Online]. Available: https://doi.org/10.1186/1476-069X-10-101

[14] S. Kamran and O. Haas, "A Multilevel Traffic Incidents Detection Approach: Identifying Traffic Patterns and Vehicle Behaviours using real-time GPS data," in 2007 IEEE Intelligent Vehicles Symposium. Istanbul, Turkey: IEEE, Jun. 2007, pp. 912–917. [Online]. Available: http://ieeexplore.ieee.org/document/4290233/

[15] T. Thianniwet and S. Phosaard, "Classification of Road Traffic Congestion Levels from GPS Data using a Decision Tree Algorithm and Sliding Windows," 2009, p. 5.

[16] Y. Ma, M. Chowdhury, A. Sadek, and M. Jeihani, "Real-Time Highway Traffic Condition Assessment Framework Using Vehicle–Infrastructure Integration (VII) With Artificial Intelligence (AI)," IEEE Transactions on Intelligent Transportation Systems, vol. 10, no. 4, Dec. 2009, pp. 615–627.

[17] F. e. a. Pedregosa, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

 ISBN: 978-1-61208-685-9