Istraživanja i projektovanja za privredu

ISSN 1451-4117

DOI:10.5937/jaes0-40553

www.engineeringscience.rs

publishing

Journal of Applied Engineering Science

Vol. 21, No. 1, 2023

Original Scientific Paper

Paper number: 21(2023)1, 1077, 346-352

DIRECTIONAL LANE CHANGE PREDICTION USING MACHINE LEARNING METHODS

Mostafa K. Ardakani [1]****, Timothy Bonds*** [2]

[1] *Information Systems and Business Analytics Kent State University, OH 44240, USA*
[2] *Senior Business Analyst, State Farm Mutual Automobile Insurance Company, Florida, USA*

** mkamalia@kent.edu*

This research employs a series of machine learning methods to predict the direction of lane change. The response is a binary variable indicating changing the lane to the left or to the right. The employed methods include Decision Tree, Discriminant Analysis, Naïve Bayes, Support Vector Machine, k-Nearest Neighbour and Ensemble. The results are compared to the conventional logistic regression method. Both performance criteria and computational times are reported for comparison purposes. A design of experiments is run to test 25 classification methods at ratios of 25%, 50%, and 75% right to left lane change data. Moreover, samples are validated by cross and holdback validation methods. RUSBoosted trees, an ensemble method, shows improvement over logistic regression. This research provides valuable insights on lane change behaviour, including trajectories and driving styles, which falls into the field of microscopic lane change study.

Keywords: lane change, decision tree, discriminant analysis, naïve bayes, support vector machine, k-nearest neighbor

1 INTRODUCTION

Intelligent mobility has changed and will continue to change the transportation industry from perspectives of policy, design, and utilization. From a public policy standpoint, interest lies in many fields such as traffic flow and safety. Sensors alert drivers to hazardous scenarios while algorithms and mathematical models seek to explain traffic phenomena. Highway settings involve two categories of models: macroscopic and microscopic. In Highway Traffic Modeling, macroscopic models address traffic issues from the perspectives of travel duration, density, speed, and flow. Microscopic models focus on the behaviors exhibited by the vehicles in traffic such as lane change and car following, Ardakani et al. (2016) [1]. Microscopic lane change models explain behaviors associated with vehicle lane changes. Lane changes fall into one of two categories, mandatory and discretionary. Mandatory lane changes consist of lane changes in which the vehicle changes to a target lane to exit or enter the highway. Discretionary lane changes involve lane changes in which the vehicle continues on the highway after arrival in the target lane. The purpose of the lane change could arise from a desire to achieve a desired speed, perception of better driving conditions in the target lane, to allow faster traffic to pass, or any other reason unassociated with mandatory lane changes, Wang (2017) [2]. For example, drivers in highway settings seek to travel at a desired speed. When the desired speed is encumbered by the speed of a leading vehicle, the driver may perform a lane change. Fig. 1 depicts the lane change of the subject vehicle *SV* from the current lane to the target lane, where $Pl_{C.}$, PV_T, and LV_T represent the preceding vehicle in the current lane, preceding vehicle in the target lane, and lagging vehicle in the target lane, respectively. The symbol *d* adjacent to the vehicle identifier (vehicle ID) denotes the space heading between vehicle *n* and the vehicle associated to the vehicle ID. More detail can also be found in Ardakani and Yang (2017) [3].

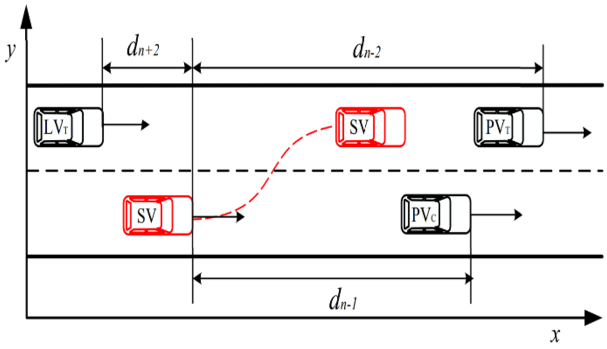

Fig 1. Lane Change to Target Lane (Gu et al., 2019) [4]

Toledo (2009) [5], built on state dependence to show state dependence along with driver heterogeneity significantly influence lane-change behavior. Sun and Elefteriadou (2012) [6] assessed a gap in lane-change trajectory models, as most of the models did not consider driver characteristics. In their research, instrument equipped vehicles monitored driver behavior during lane-change scenarios. Using K-means clustering, divided the drivers into four categories. Ekbatani et al. (2016) [7] found heterogeneity in drivers led to divergent strategies for lane choice and lane choice was subject to the speed choice of the driver. Wang (2017) [2] addressed heterogeneity in driver behavior through a focus on aggressive lane-change behavior. Aggressive lane-change behavior includes numerous lane changes caused by weaving in and out of traffic, overtaking a vehicle with a lane change to the immediate right of the slower vehicle, and lane changes without signaling. With a focus on aggressive lane-change behavior, Wang (2017) [2] created a model for driving simulation training. He also showed as traffic flow decreases, the percentage of aggressive drivers increase, and the aggregate number of lane changes increase. Ren et al. (2019) [8] proposed a new free lane changing model based on machine learning, which considered three driving styles, cautious, stable, and radical. Using regression to smooth NGSIM data, then affinity propagation clustering, Pang (2019) [9] developed a model to account for differing driver characteristics. De Zepeda et al. (2021) [10] stated driving styles are influenced by a combination of human and environmental factors. As environmental factors change, the driver adapts the desired driving style accordingly. As such, potential exists for a driver to display atypical lane change behavior and become classified in alignment with such behavior due to unfavorable environmental influences. Additionally, they referenced the noteworthy insights gleaned from prior research on lane change models with consideration of driving style. They also noted driving styles, such as aggressive, normal, and passive, may not reflect complicated behaviors in real world settings. Also, more recent works that are related to lane change and machine learning topics can be found at Yang et al. (2022a) [11], Yang et al. (2022b) [12] and Kruger et al. (2019) [13].

The focus of this research is to predict lane change as a binary choice after the driver has committed to the execution of a lane change, the left or right lane. In addition to conventional logistic regression, decision tree, discriminant analysis, naïve bayes, support vector machine, k-nearest neighbor and ensemble are assessed. The main contribution of this research is to compare existing machine learning methods and discuss their merits and disadvantages. The paper is organized as follows. Section 2 describes data and data structure. Section 3 outlines applied methodology. The results are provided in Section 4. Concluding remarks are provided in the final section.

2 DATA

The Next Generation Simulation (NGSIM) [14] program is a collection of vehicle trajectory data from US 101 and Lankershim Boulevard in Los Angeles, Interstate-80 in Emeryville, CA, and Peachtree Street in Atlanta. The column headings within the NGSIM dataset are shown in Table I. Each row of data for a specific vehicle ID represents a decisecond interval. The following NGSIM fields used for this analysis includes velocity and space heading, duration, and density.

V = average velocity of vehicle *n*

gn-1 = space heading (distance) between vehicles *n* and *n-1* before the lane change

T = duration of the lane change measured in seconds

D = vehicles per kilometer

A total of 700 lane changes were randomly selected from NGSIM US 101 and Interstate-80 to analyze discretionary highway lane changes. Data pre-processing includes the removal of mandatory lane changes from lane seven. Moreover, data cleaning is performed. Some cases are also excluded from our data to represent more realistic scenarios. For instance, when a vehicle within the NGSIM dataset does not have leading vehicle, the space heading noted is 0. The remaining lane changes consisted of 513 left and 157 right lane changes. Table II contains a preview of the analyzed dataset.

Table 1. Partial Data Points from the NGSIM Dataset

Row Number	Data Analyzed				
	V (m/s)	*gn-1 (m)*	*T (s)*	*D (vehicles/km)*	*Lane Choice (0 or 1)*
1	7.167	30.937	2	25.76	1
2	10.500	20.525	1.7	62.69	0
3	10.490	20.803	7.1	31.44	1
4	6.628	40.233	3.8	40.28	1
5	9.402	12.405	3	103.6	0
6	9.123	25.408	4.9	94.89	1
7	16.726	17.401	1.7	17.06	0
8	10.527	27.408	28.8	85.606	1
9	10.784	24.844	2.2	32.386	1
10	5.831	24.283	5	40.72	1
…	…	…	…	…	…

Table 2. Data Preview

NGSIM Metrics	18 NGSIM Data Points	
	Column Name	*Data Type*
1	Vehicle ID	Int
2	Frane ID	Int
3	Total Frames	Int
4	Global Time	Int
5	Local X	Double
6	Local Y	Double
7	Global X	Double
8	Global Y	Double
9	Vehicle Length	Double
10	Vehicle Width	Double
11	Vehicle Class	Categorical
12	Vehicle Velocity (ft/sec)	Double
13	Vehicle Acceleration (ft/sec)	Double
14	Lane ID	Int
15	Preceding Vehicle ID	Int
16	Following Vehicle ID	Int
17	Space Headway	Double
18	Time Headway	Double

3 METHODOLOGY

As noted by Wang (2017) [2], desired speed may be one reason to explain a lane change; however, the exact reason for a right and left lane change may differ. For example, exit lanes traffic are typically on the right. This analysis does not include mandatory lane change exits from the highway; however, the dataset does include right lane changes which preceded the mandatory lane change exit. For this reason and the small sample of right lane changes, a design of experiments was conducted to determine the best mix of left and right lane changes. Three levels are considered for the percentage of left or right samples, namely, 25%, 50% and 75%. For details about design of experiments, see Ardakani (2016) [15] and Ardakani and Wulff (2013) [16]. For each experiment, the sample of lane changes were randomly selected to achieve the respective ratio of right-to-left lane changes. The design of experiments also incorporated two methods of validation, cross and holdback.

In cross validation, a user defines a specific number of folds (*k*). The dataset is then split into k randomly selected folds, with *k-1* folds used as the training set. The remaining fold is reserved as the testing set. This process is repeated *k* times, and the overall accuracy is an average of all iterations. Holdback validation requests a user defined percentage of data to be set aside as a test set. Subsequently, the classification method trains a model using the training set and assesses performance with the test set; see Cross-Validation (2021) [17]. The full data set is used to train the final model. MATLAB 2020a software is used for all analyses. Also, readers can find details about all discussed methods and commands in Statistics and Machine Learning Toolbox manual; the MathWorks, Inc. (2020) [18]. The CPU for the computer on which the analyses was run is an Intel Core i5-8250U, with a base frequency of 1.6 GHz and a maximum turbo frequency of 3.4 GHz.

4 RESULTS

Each experiment was conducted against 25 classification methods with 8 methods run at a time in parallel. Table III includes a list of all 25 classification methods along with their accuracy and performance results. Classical statistics promotes the use of logistic regression as an acceptable method to model the pattern of binary dependent variables, Bera et al. (2020) [19]. This method is also acceptable for prediction, Rahman et al. (2021) [20]. Therefore, the results from the logistic regression classification method are considered the benchmark and are compared to the results from the most accurate method. The experiment with a mixture of 75% right lane changes, 20% holdback validation and RUSBoosted trees ensemble method produced the highest accuracy at 90.2%. The mixture of the analyzed dataset lends itself to the benefits of RUSBoosted trees. This method efficiently classifies imbalanced training data through data sampling and boosting. Class imbalance occurs when one class in the dataset greatly outnumbers the other class(es). Typically, data mining algorithms encounter difficulties producing optimal models with such data, Seiffert et al. (2009) [21]. Table IV shows the confusion matrix for this RUSBoosted trees. Table V shows the confusion matrix for logistic regression. Logistic regression achieved 80.5% accuracy for the same experiment while this value for RUSBoosted method is 90.2%. To compare the performance of the two methods, Table VI shows the prediction speed and training time for all methods in addition to performance criteria including accuracy, sensitivity, positive predictive (precision) value, negative predictive value, and F1 score. In overall, we can conclude that RUSBoosted performs better than classical logistic regression. Figure 2 shows that RUSBoosted Trees ensemble method is an improvement over the benchmarked logistic regression method in terms of accuracy. However, logistic regression is faster in both training time and prediction speed as shown in Figures 3 and 4.

Table 3. List of 25 Classification Methods using Design of Experiments with a 75% Right to Left Lane Ratio and 20% Holdback Validation

Method Category	25 Classification Methods			
	Method	*Accuracy*	*Training Time (seconds)*	*Prediction Speed (observations/second)*
Decision Tree	Fine Tree	78%	1.985	1600
Decision Tree	Medium Tree	78%	1.3959	1600
Decision Tree	Coarse Tree	70.7%	1.0473	15000
Discriminant Analysis	Linear Discriminant	75.6%	0.69496	13000
Discriminant Analysis	Quadratic Discriminant	63.4%	0.74761	12000
Regression	Logistic Regression	80.5%	0.34013	8900
Naïve Bayes	Gaussian Naïve Bayes	65.9%	0.7175	15000
Naïve Bayes	Kernel Naïve Bayes	76.6%	0.89327	5100
SVM	Linear SVM	75.6%	0.76359	15000
SVM	Quadratic SVM	75.6%	0.64341	12000
SVM	Cubic SVM	70.7%	0.53002	14000
SVM	Fine Gaussian SVM	73.2%	0.41703	12000
SVM	Medium Gaussian SVM	75.6%	0.38895	1200
SVM	Coarse Gaussian SVM	75.6%	0.28934	13000
KNN	Fine KNN	78%	0.16585	14000
KNN	Medium KNN	75.6%	0.032137	7200
KNN	Coarse KNN	75.6%	1.0655	9100
KNN	Cosine KNN	75.6%	0.92045	9200
KNN	Cubic KNN	75.6%	0.75007	7200
KNN	Weighted KNN	75.6%	0.6079	9700
Ensemble	Boosted Trees	80.5%	1.6919	1200
Ensemble	Bagged Trees	87.8%	1.083	1100
Ensemble	Subspace Discriminant	75.6%	0.93045	920
Ensemble	Subspace KNN	73.2%	1.4703	650
Ensemble	RUSBoosted Trees	90.2%	1.0239	1200

Table 4. Confusion Matrix for the RUSBoosted Trees

True Values	Confusion Matrix for RUSBoosted Trees	
	Predicted 0	*Predicted 1*
True 0	9	1
True 1	3	28

Table 5. Confusion Matrix for Logistic Regression

True Values	Confusion Matrix for Logistic Regression	
	Predicted 0	*Predicted 1*
True 0	3	7
True 1	1	30

Table 6. Accuracy and Performance Comparison

Classification Methods	Accuracy and Performance Results						
	Accuracy (%)	*Sensitivity (%)*	*Precision-positive predictive value (%)*	*Negative predictive value (%)*	*F1 score (%)*	*Training Time (seconds)*	*Prediction Time (observations per second)*
RUSBoosted Trees Ensemble	90.2	90	75	96.5	82	1.0239	1200
Logistic Regression	80.5	30	75	81	42	0.34013	8900

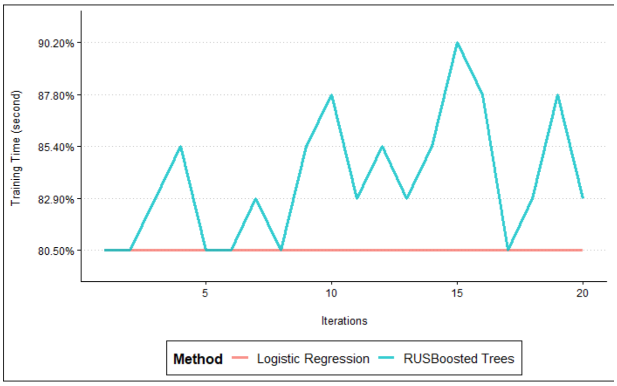

Fig 2. Accuracy Comparison

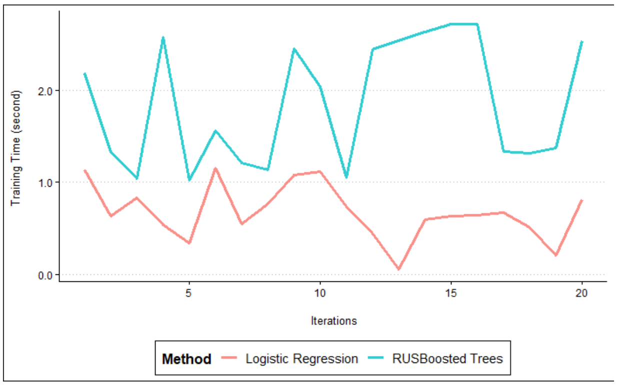

Fig 3. Training Time Comparison

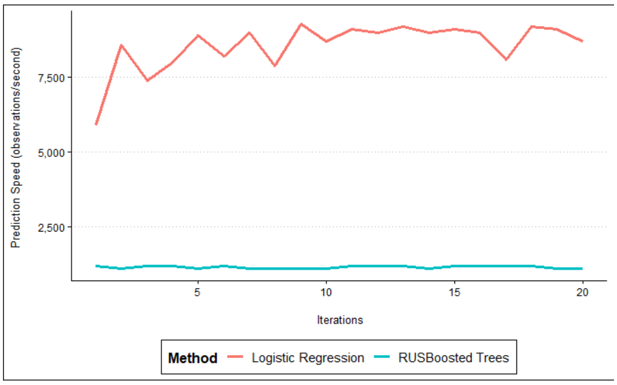

Fig 4. Prediction Speed Comparison

5 CONCLUSION

This research tackled an important problem, which has applicability to the autonomous vehicle and traffic simulation industries. In this study, machine learning methods are proposed as an alternative to classical logistic regression to predict the direction of a lane change. NGSIM dataset is used to generate a lane change dataset. Also, data cleaning was performed to remove outliers and noises from data. Through a design of experiments, 25 classification methods were tested against randomly selected containing ratios of 25%, 50%, and 75% right to left lane change data. Subsequently, each sample was validated through both cross and holdback validation methods. RUSBoosted trees, an ensemble method, achieved a 12% improvement in accuracy over logistic regression. For RUSBoosted method, performance criteria in percentages for accuracy, sensitivity, positive predictive, negative predictive, and F1 score are 90.2, 90, 75, 96.5, and 82, respectively; while these values for logistic regression are 80.5, 30, 75, 81, 42. Therefore, with respect to performance, RUSBoosted trees is a suitable alternative to logistic regression. It should be noted that logistic regression remains a superior method if we only take training time or prediction speed into consideration. Future research will expand on results achieved through the addition of data for vehicles in adjacent lanes. The desire is incorporate aspects of traditional lane change models to predict the direction of lane change given space headings and velocities leading and following vehicles in both lanes. Upon achieved success, the model can be tested with traffic simulation software and further optimized to enable autonomous vehicles to continually calculate the ideal target lane, left or right, based on the traffic the given traffic scenario in adjacent lanes.

6 REFERENCES

[1] Ardakani, M. K., and J. Yang. (2016). Generalized Gipps-type vehicle-following models. Journal of Transportation Engineering Part A, 143 (3): 04016011.https://doi.org/10.1061/JTEPBS.0000022

[2] Wang, Y. (2017). Modeling and Simulation of Aggressive Lane-Changing Behavior for Highway Driver Training. 2017 3rd IEEE Internaitonal Conference on Computer and Communications (pp. 2894-2898). Chengdu: IEEE

[3] Ardakani, M. K., Yang, j., and Sun, L., (2017). Stimulus response driving behavior: An improved general motor vehicle-following model, Advances in Transportation Studies, no. 39, pp. 23–36

[4] Gu, X., Yu, J., Han, Y., Han, M., & Wei, L. (2019). Vehicle Lane Chnage Decision Model Based on Random Forest. 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), (pp. 115-120). Shenyang, China

[5] Toledo, T. a. (2009). State Dependence in Lane-Changing Models. Transportation Research Record: Journal of the Transporation Research Board, 2124(1), 81-88

[6] Sun, D. a., and Elefteriadou L., (2012). Lane-Changing Behavior on Urban Streets: An “In-Vehicle” Field Experiment-Based Study. Computer-Aided Civil and Infrastructure Engineering, 27(7), 525-542

[7] Ekbatani Keyvan, M. V. (2016). Categorization of the Lane Change Decision Process on Freeways, Transportation Research Part C: Emerging Technologies, 69, 515-526

[8] Ren, G., Zhang, Y., Liu, H., Zhang, K., & Hu, Y. (2019). A New Lane-Changing Model with Consideration of Driving Style. International Journal of Intelligent Transportation Systems Research, 17, 181-189

[9] Pang, M.-Y. (2019). Trajectory Data Based Clustering and Feature Analysis of Vehicle Lane-Changing Behavior. 2019 4th International Conference on Electromechanical Control Technology and Transportation (ICECTT) (pp. 229-233). Guilin: IEEE

[10] De Zepeda, M. V., Meng, F., Su, J., Zeng, X.-J., & Wang, Q. (2021). Dynamic Clustering Analysis for Driving Styles Identification. Engineering Applications of Artificial Intelligence, 97

[11] Yang Yang, Kun He, Yun-peng Wang, Zhen-zhou Yuan, Yong-hao Yin, Man-ze Guo, (2022a). Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods, Physica A: Statistical Mechanics and its Applications, Volume 595, 127083

[12] Yang Yang, Kun Wang, Zhenzhou Yuan, Dan Liu, (2022b). "Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction", Journal of Advanced Transportation, vol. 2022, Article ID 4257865, 16, https://doi.org/10.1155/2022/4257865

[13] Krüger, Martin, Anne Stockem Novo, Till Nattermann, Torsten Bertram, (2019). Probabilistic Lane Change Prediction using Gaussian Process Neural Networks. In 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019, Auckland, New Zealand, October 27-30, 2019. pages 3651-3656, IEEE

[14] Next Generation Simulation (NGSIM). (2020, November 2). Retrieved from Federal Highway Administration: https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm

[15] Ardakani, M.K., (2016). The impacts of errors in factor levels on robust parameter design optimization, Quality and Reliability Engineering International 32 (5), 1929-1944

[16] Ardakani, M.K., Wulff S.S., (2013). An overview of optimization formulations for multiresponse surface problems, Quality and Reliability Engineering International 29 (1), 3-16

[17] Cross-Validation. (2021). Retrieved from Mathworks(r): https://www.mathworks.com/discovery/cross-validation.html#:~:text=Cross-Validation%20with%20MATLAB%20MATLAB%20%C2%AE%20supports%20cross-validation%20and,app%20for%20training%2C%20validating%2C%20and%20tuning%20classification%20models

[18] The MathWorks, Inc. (2020). MATLAB Statistics and Machine Learning Toolbox User Guide. Natick, MA: The MathWorks, Inc.

[19] Bera, B., Saha, S., & Bhattacharjee, S. (2020). Forect Cover Dynamics (1998 to 2019) and Prediction of Deforestation Probability using Binary Logistic Regression Model of Silabati Watershed, India. Trees, Forests and People, 2

[20] Rahman, H. A., Wah, Y. B., & Huat, O. S. (2021). Predictive Performance of Logstic Regression for Imbalanced Data with Categorical Covariate. Pertanika Journal of Science & Technology, 29(1), 181-197

[21] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Transacation on Systems, Man, and Cybernetics, 40(1), 185-197

Paper submitted: 08.10.2022.

Paper accepted: 22.12.2022.