Noname manuscript No.
(will be inserted by the editor)

Driver Behavior Profiling using Machine Learning

Soumajit Mullick · Pabitra Mohan Khilar

Received: date / Accepted: date

Abstract The drivers' behavior influences the traffic on road and this, in turn influences energy consumed by the vehicles and emission of pollutants from the vehicles. So it is necessary to identify drivers' characteristics to profile their behavior correctly. A large amount of data is needed for the analysis which is collected by the on-Board Unit present on the vehicle. On-Board Unit has sensors that are used to collect the required data. The comparative performance of different machine learning algorithms is evaluated on the data collected by the on-board unit and in turn help in profiling drivers' behavior. The experimental result shows that the support vector machine gives an accuracy of 99.4% amongst the remaining classifier.

Keywords Vehicular Ad hoc Networks (VANET) · Machine Learning · Intelligent Transportation Systems (ITSs) · Driver Behaviour · Safety Application

1 Introduction

Nowadays, the number of vehicles on the road is increasing exponentially. The population of considerable size uses a private car for their daily commute. One of the major drawbacks of using such a huge number of vehicles is road accidents. This leads to traffic jams and in turn leads to high fuel consumption and emission of dangerous pollutants from the vehicles. Dangers and expenses linked to road accidents are treated as a serious problem in today's society. The statistic related to the number of accidents on an Indian road is released by the government, which is alarming. In 2017, the total road accident was reported to be 4,64,910, which claimed 1,47,913 lives and 4,70,975 persons were injured. This can be interpreted

Soumajit Mullick
Department of Computer Science and Engineering
National Institute of Technology, Rourkela
E-mail: 218CS1093@nitrkl.ac.in

Pabitra Mohan Khilar
Department of Computer Science and Engineering
National Institute of Technology, Rourkela

into 1,290 injured people and 405 lives lost daily from 1,274 accidents. The fact is alarming that this is the official number and does not include the accidents which were not reported. [1]

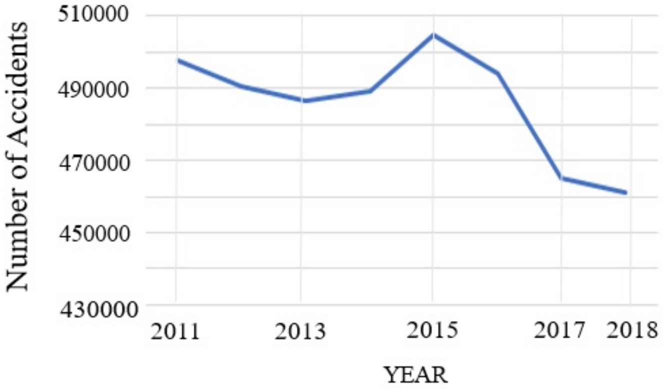


Fig. 1: Road side accident statistic in India

According to the National Highway Traffic Safety Administration (NHTSA) about 25% of police reported crashes involve some form of driver inattention. One of the major reasons for the road accidents is the careless nature of a driver. The carelessness not only hurt himself but also the other people who are riding with him and also all other vehicles on the roads. This cause major troll on the number of families and this happens only due to the negligence (known or unknown to himself) of some drivers. Therefore, emerging technologies to sense and alert oblivious drivers are very important, to avert vehicular mishap and to imbibe disciplined driving in the drivers.

Advancement of wireless communication, which is applied to mobile computing, has boosted the intelligent transportation system(ITS) where the main focus is on the development of road safety applications[2] [3]. The technology that binds the above wireless communication with the automobile industry to take the industry to the next level is VANET. This is the backbone of Intelligent Transportation Systems (ITS)[4]. VANET establishes a connection between vehicles(V2V) as well as vehicles and the roadside unit(RSU) (V2I). As it is an ad-hoc network it does not need much infrastructure to build the network. The presence of a communication unit (On-board Unit) helps to use VANET for several applications like convenience applications, productive applications, commercial applications. Convenience applications include toll tax collection[5], automatic parking service[6], etc. Productive applications include environmental parameter monitoring[7], secure transaction through VANET[8], etc. The commercial application includes marketing on the wheel[9]. Various optimization techniques to optimize the generic parameter of VANET [10] during application makes VANET popular among researchers.

In VANET, nodes communicate with each other using short-range wireless communication (e.g., IEEE 802.11p). A special allocation of 75 MHz in the 5.9 GHz is done by the FCC(Federal Communications Commission) for licensed DSRC

(Dedicated Short Range Communication) for the communication between vehicles and infrastructure whose main focus is to improve bandwidth utilization and to bring down latency.

The remainder of this paper is explained as follows. In section 2, we have discussed some previous work done in this field. Section 3, we discussed various machine learning techniques we have used. Section 4, we described the dataset used, preprocessing and implementation of the algorithms. Section 5, we discussed the experiment result and finally we conclude the paper and the future work which can be done.

2 Related Work

Since past few years, there are research being conducted going on monitoring driver behavior and also automatic road accident detection using various methodologies. Many researchers have ventured into the measuring of fatigued and drunkenness of the driver and also various risky behaviors which may be prone to accident. We have to take into consideration that there are many live applications related to insurance domain and fleet management. But they are not publicly accessible to us for research work. Some examples include Aviva Drive, Ingenie, Greenroad, Snapshot, and Seeing Machines.

Nericell, proposed by Mohan et al.[11], is an application based on Windows to monitor road traffic and the condition of the road. An accelerometer is used to detect potholes and braking events. It also uses GPS/global system of mobile (GSM) communications to obtain the locality of the vehicle. Braking, bumps, and potholes are detected using a predefined threshold value. No machine learning algorithms were employed to learn the threshold value. Detection results in terms of False Positives (FPs) and False Negatives (FNs) are given in the table .

Table 1: Nericell application result

Event	**False Negetive**	**FalsePositive**
Braking events detection	4.4%	22.2%
Bumps/potholes detection	23%	5%
Honk detection	0	0

The android application [12] was proposed by Dai and colleagues for real time detection of dangerous driving events and alert the driver. This detection is related to DUI of alcohol. Smartphone orientation and accelerometer are to detect Abnormal Curvilinear Movements (ACM) and Problems in Maintaining Speed (PMS), which in turn, related to the detection of drunk driving behavior.

MIROAD is an iPhone based application created by Johnson and Trivedi[13]. It uses a magnetometer, accelerometer, gyroscope, and GPS sensor data from the smartphone to classify whether the driver is aggressive/non-aggressive. It uses the DTW algorithm processed in real-time on the smartphone. The experiment result shows 97% accuracy.

SenseFleet is an application based on the Android platform devices, proposed by Castignani et al. [14], to detect unsafe driving events which are independent of vehicle and mobile device. It collects data from the magnetometer, gravity sensor, accelerometer and GPS from the mobile device. This data is used in a fuzzy system detection for risky behavior.

In driver behavior profiling, Onboard Unit(OBU) is fixed with a sensor like a gyroscope, accelerometer, GPS to collect data. OBU is kept in the vehicle. To collect data, a driver simulates a real-life risky behavior scenario. This is then modeled to classify genuine driving and help to profile the driver. In some papers, various sensors like alcohol sensor, eye movement, eye exposure are used to detect the health of the driver.

3 Machine Learning Algorithm

Here, we have used three machine learning algorithms. They are described below:

3.1 Logistic Regression

Logistic regression [15] is a supervised learning model use for the classification of non-linear data borrowed from the statistic field. Here, the target value can only be discrete values. It learns to calculate the probability of a given data consisting of a number of features belonging to a particular class. It uses the sigmoid function ($g(z) = 1/(1 + e^{-z})$) to calculate the probability.

3.2 SVM

SVM(Support Vector Machine) [16] is a machine learning model for classification as well as regression of data. It is a supervised learning algorithm that learns to find an optimal hyperplane that maximizes the separation (distance between the margin) of the training data. These margins are known as support vectors.

3.3 Multi Layer Perceptron

Multi-Layer Perceptron is a simple 2 layer Artificial Neural Network (ANN). It consists of one input layer, one hidden layer, and one output layer. Hidden layer and output layer consist of nodes that are connected using different weights (these weights are learned by the algorithm using training data). The human brain was the inspiration behind the ANN algorithm to learn complex data[17]. Number of nodes in input layer is equal to the number of features and number of nodes in the output layer is equal to the number of classes model will classify into. A single node computes the sigmoidal transfer of a weighted sum of value from the output of the previous layer.

4 Methodology

For the paper, we have used the driver behavior dataset created by [18]. The dataset is a collection of sensor's reading placed on the vehicle. The sensors used were accelerometer and gyroscope. The experiment was done with the help of three drivers to reproduce real like events on the road. The experiment's condition are as follow:

- Cars : Ford Fiesta 1.25, Ford Fiesta 1.4, Hyundai i20
- Driver Population : Three drivers of age 26, 27, 28
- Driver Behaviors : Sudden Right Turn, Sudden Break, Sudden Left Turn, Sudden Acceleration
- Sensor used : MPU6050
- Device used: Raspberry Pi 3 Model B

The purpose of this dataset was to record a set of driving events that represented real-world driving behaviour such as braking, accelerating, turning, and lane changes. The raw data is stored in a CSV file in row format as "GyroX GyroY GyroZ AccX AccY AccZ", where (GyroX, GyroY, GyroZ) is 3d coordinate of gyroscope and (AccX AccY AccZ) is 3d coordinate of an accelerometer. 15 consecutive rows are grouped into the temporary sliding window as shown in Figure 2.

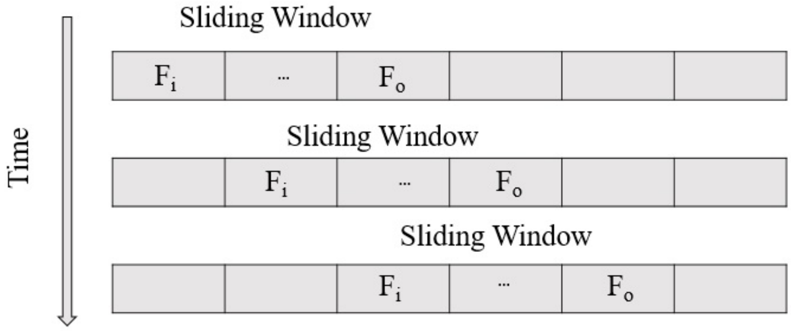


Fig. 2: Sliding window mechanism

This window is moved by one row at a time. For each window, the following statistics are calculated:

1. Sum : $\sum_{i=1}^{N} x$
2. Mean : $\frac{\sum_{i=1}^{N} x}{N}$
3. Median : $med(X)$
4. Variance : $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
5. Min : $Min(X)$
6. Max : $Max(X)$
7. Skewness : $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$

8. Kurtosis : $\frac{\frac{1}{N}\sum\limits_{i=1}^{N}(x_i-\bar{x})^4}{s^4}$

From timestamp dependent data in the raw dataset, using the above sliding window we have transformed into independent data. We have used 8 statistic measures. As a result, the number of features of the processed dataset is 48 i.e. 8 * 3 * 2 (8 statistic measures, 3 axes of a sensor, 2 sensors). Data on the above dataset is not normalized. To increase the performance of the model, we have normalized the above data.

5 Result Analysis and Discussion

We applied three machine learning algorithms: Logistic regression, SVM and MLP on the processed independent data. We divided the data into training, cross validation and testing in 3:1:1 ratio. We have used the following measures to compare the performance of the above algorithms:

- Accuracy
- Precision
- Recall
- F1 score

Results of the experiment is shown in following tables:

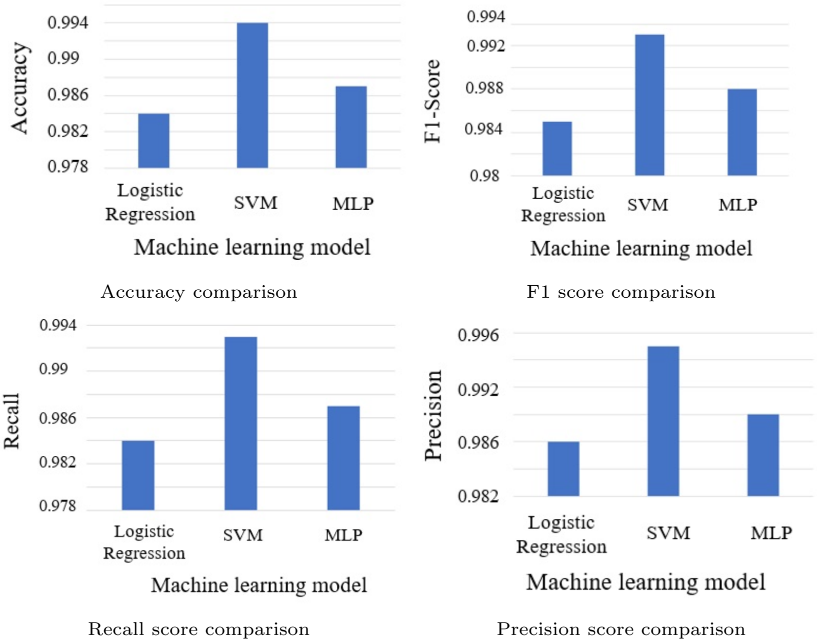
Table 2: Experiment Result of 3 MLAs without Normalization

MLA	**Accuracy**	**Precission**	**Recall**	**F1 score**
Logistic Regression	0.963	0.967	0.961	0.964
SVM	**0.984**	**0.989**	**0.988**	**0.988**
MLP	0.971	0.973	0.975	0.974

Table 3: Experiment Result of 3 MLAs with Normalization

Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 score**
Logistic Regression	0.984	0.986	0.984	0.985
SVM	**0.994**	**0.993**	**0.993**	**0.993**
MLP	0.987	0.988	0.987	0.988

From the above result we can see that support vector machine perform best with accuracy close to 99.4%. SVM tries to separate the two-class such that the distance between the two margins is maximal. So it will find a solution that is as reasonable as possible for both groups. This property does not hold by both linear regression and multi-layer perceptron. This is the reason for support vector machine's performance is better than both the models.



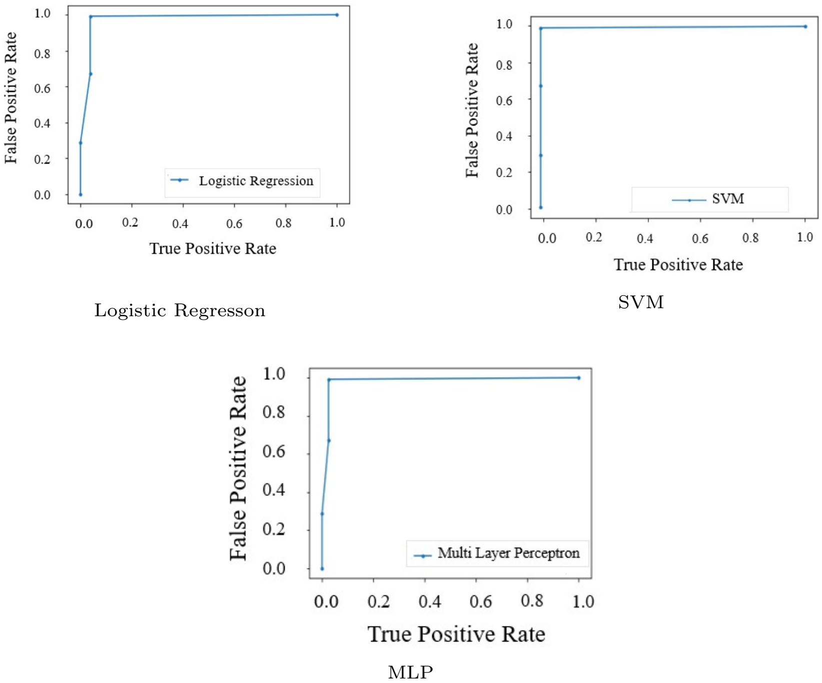
Accuracy comparison

F1 score comparison

Recall score comparison

Precision score comparison

Fig. 4: Metrics comparison of ML model



Logistic Regresson

SVM

MLP

Fig. 6: ROC of machine learning models

6 CONCLUSION AND FUTURE WORK

We have shown in this paper, the comparative study of three machine learning algorithms: Logistic regression, SVM and MLP, using available dataset. From the evaluation matrix it can be observed that SVM performs better than other two MLAs. We can conclude that 3 axes data from the 2 sensors are necessary for classification. Normalization of data is needed to increase the performance of all the machine learning model. All 8 statistic measures were critical in increasing the accuracy. For future work, we want to install the sensor devices on all vehicles present in a particular town to get a real dataset with various weather and road conditions. We can use this large dataset to model a deep learning algorithm such as Long Short-Term Memory(LSTM) network to get better results and gain new insight into driver behavior detection.

References

1. Road accidents in india claim more than 14 lakh lives in 2017. https://www.autocarindia.com/industry/road-accidents-in-india-claim-more-than-14-lakh-lives-in-2017-410111
2. S. Olariu, M.C. Weigle, *Vehicular networks: from theory to practice* (Chapman and Hall/CRC, 2009)
3. Y. Qian, N. Moayeri, in *VTC spring* (2008), pp. 2794–2799
4. S.K. Bhoi, P.M. Khilar, IET networks **3**(3), 204 (2013)
5. B.R. Senapati, P.M. Khilar, N.K. Sabat, in *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)* (IEEE, 2019), pp. 1–5
6. H. Zhao, L. Lu, C. Song, Y. Wu, International Journal of Distributed Sensor Networks **8**(12), 280515 (2012)
7. B.R. Senapati, R.R. Swain, P.M. Khilar, in *Smart intelligent computing and applications* (Springer, 2020), pp. 229–238
8. K.E. Shin, H.K. Choi, J. Jeong, in *Proceedings of the 4th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks* (ACM, 2009), pp. 175–182
9. S.K. Bhoi, D. Puthal, P.M. Khilar, J.J. Rodrigues, S.K. Panda, L.T. Yang, Computer Networks **142**, 168 (2018)
10. B.R. Senapati, P.M. Khilar, in *Nature Inspired Computing for Data Science* (Springer, 2020), pp. 83–107
11. P. Mohan, V.N. Padmanabhan, R. Ramjee, in *Proceedings of the 6th ACM conference on Embedded network sensor systems* (ACM, 2008), pp. 323–336
12. J. Dai, J. Teng, X. Bai, Z. Shen, D. Xuan, in *2010 4th International Conference on Pervasive Computing Technologies for Healthcare* (IEEE, 2010), pp. 1–8
13. D.A. Johnson, M.M. Trivedi, in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (IEEE, 2011), pp. 1609–1615
14. G. Castignani, T. Derrmann, R. Frank, T. Engel, IEEE Intelligent Transportation Systems Magazine **7**(1), 91 (2015)
15. D.G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, *Logistic regression* (Springer, 2002)
16. V. Vapnik, *The nature of statistical learning theory* (Springer science & business media, 2013)
17. C. Zhao, Y. Gao, J. He, J. Lian, Engineering Applications of Artificial Intelligence **25**(8), 1677 (2012)
18. A. Asim Sinan Yuksel. Driving behavior dataset. http://dx.doi.org/10.17632/jj3tw8kj6h.1file-83a10979-d980-4099-b63f-d3e6f809d8e3