# Apply Scikit-Learn in Python to Analyze Driver Behavior Based on OBD Data

[+]Chi-Pan Hwang [++]Mu-Song Chen [*]Chih-Min Shih [**]Hsing-Yu Chen [***]Wen Kai Liu
[+]Department of Electronic, National Changhua University of Education
cphwang@cc.ncue.edu.tw
[++]Department of Electric, Da-Yeh University
chenms@mail.dyu.edu.tw
Institute of Information Industry
[*]danielshih@iii.org.tw [**]edwardchen@iii.org.tw [***]wkliu@iii.org.tw

***Abstract*—The long term accumulated driving information can effectively summarize the specific driver behavior by statistical analysis. In order to widely and chronically collect driving information of drivers, the cloud computing platform is the most suitable mechanism to log the dynamic vehicle information stream from OBD port to build up Big Data for data mining about driver behavior, currently. The research of this paper has focused on the application layer in the cloud computing platform, Python has been adopted to as the main development tool accompanying with the packages of *numpy*, *pandas*, and *scipy* to calculate the kurtosis and skewness in statistics of each driving route, then decision tree classification technique was applied to generate the analyzing knowledge for driver behavior analysis. Finally the driver behavior are summarized from the completed decision tree classifier to defensive, weak defensive, weak aggressive, and aggressive to complete the overall operations.**

***Keywords—Data mining, Driver Behavior, Python, Scikit-learn***

## I. Introduction

The incidence of vehicle accidents is closely related to driving behavior. Most of driver behavior was evaluated by the designated index of driving habit. The driving habit index evaluation of a driver must be followed by a long term period. The long term accumulated driving information can effectively summarize the specific driver behavior by statistical analysis. In order to widely and chronically collect driving information of drivers, the cloud computing platform is the most suitable mechanism to log the dynamic vehicle information stream from OBD(On-Board Diagnosis) port to build up Big Data for data mining about driver behavior, currently [1][2].

The research of this paper has focused on the application layer in the cloud computing platform, Python has been adopted to as the main development tool accompanying with the Scikit-learn [3] packages of *numpy* (scientific computing), *sklearn* (machine learning), and *scipy* (science and statistical computing) to calculate the kurtosis and skewness in statistics of each driving route, then decision tree classification technique was applied to generate the analyzing knowledge for driver behavior analysis.

The entire analysis process is divided into iterative processes of *data clean*, *data integration*, *data selection*, *data transformation*, *data mining*, as well as *knowledge evaluation/display* to achieve [4], in which data mining and knowledge evaluation/display are the core of these analysis processes. In the data mining process phase, the vehicle speed and acceleration of each logged route are used as input parameters to calculate the statistical kurtosis and skewness which further apply to establish decision tree as the qualitative analysis results of driving behavior. The knowledge evaluation/display phase mainly summarizes the classification rules from the completed decision tree classifier, as well as interpret driver behavior category semantics to defensive, weak defensive, weak aggressive, and aggressive to complete the overall processes of driver behavior.

The paper is organized as follows. Section 2 illustrates the system architecture which includes the function descriptions. Section 3 describes the implementation of system function and application software systems, verifies the system functions, and exemplifies the application operations. Section 4 concludes the development results and further study topics.

## II. System Abstraction Architecture

Fig. 1 depicts an abstract system architecture being composed by cloud computing platform, web server, and data processing unit. The huge amount of driving dynamics data have been collected by using cloud computing technology on the cloud computing platform that includes the GPS and OBD datasets. The GPS dataset has GPS velocity, position, and time items. The OBD dataset contains OBD velocity, engine RPM, intake air flow rate, engine coolant temperature, engine load, fuel consumption rate, and battery voltage items. The web server is an interface between data processing unit and cloud computing platform. It provides web API (Application Programming Interface) for data processing unit to retrieve the stored driving dynamics data from cloud computing platform. The data processing unit designed to have functions of driver behavior analysis, road traffic monitor, and vehicle health care where driver behavior analysis is the main theme in this paper. Finally, the processing results are returned to cloud computing platform to support the development of other related applications.

978-1-5386-5395-1/18/$31.00 ©2018 IEEE
DOI 10.1109/WAINA.2018.00159

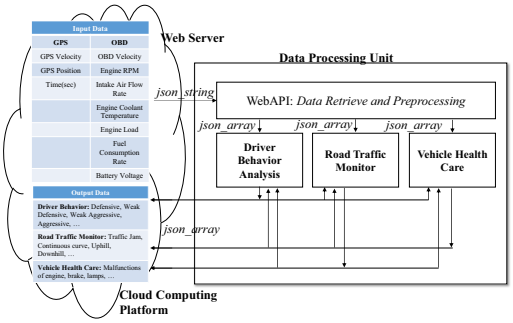



Fig. 1 Abstract system architecture of driver behavior analysis

## III. System Implementation

The driver behavior analysis consists of the sequence processes of *data clean*, *data integration*, *data selection*, *data transformation*, *data mining*, and *knowledge evaluation/display* as shown in Fig. 2. They are implemented by Python programming language based on Scikit-learn class packages including numpy, scipy, pandas, etc.,. The data clean process retrieves a logged route data from cloud computing platform as input using web API with an authorized user key and route key. It returns a JSON string list that contains JSON strings of RM (Return Message), RC (Return Code), AvgFuel (Average Fuel Consumption), and Distance as well as a JSON object of DATA. In which, the DATA object is the input of data clean process that contains a logged route datasets of GPS and OBD. It is further partitioned into individual JSON strings of vehicle dynamic parameters and converted to numerical arrays which are collected into $\hat{\boldsymbol{R}}_k$ by Excel format.

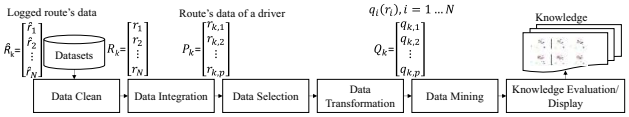
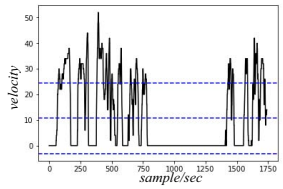


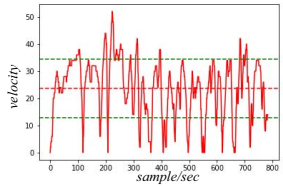

Fig. 2 Processes of driver behavior analysis

### 3.1 Data Clean

The flaws detection of retrieved logged route data was an important work of data clean process implementation. A velocity of logged route data was selected to calculate the mean value and standard deviation. The result was shown in Fig. 3(a). The calculated mean value is 10.69, and standard deviation is 13.82. The estimated minimum velocity is less than zero. It is not reasonable. Because there many zero velocities are found that were caused by suspend for traffic jam or sign. Therefore, it is a necessary process to remove redundant zero velocities of logged route data in data clean. Fig. 3(b) illustrates the result of redundant zero velocities being removed that calculated mean value is 23.71, and standard deviation is 10.73. The estimated velocities were in reasonable range. Consequently, these redundant zeros of a logged route data $\hat{\boldsymbol{R}}_k$ must be removed to generate $\boldsymbol{R}_k$ before that enters to subsequent processes. Beside of that, the Kalman filter has been implemented in data clean process to filter the noises caused by data sampling and communication.




(a) With redundant zero velocities




(b) Without redundant zero velocities

Fig. 3 Flaws detection of redundant zero velocities

### 3.2 Data Integration

The cleaned R is in Excel file format. Each row $r_i$ represents a parameter of a logged route dataset, for instance velocity, engine RPM, or other parameters mentioned in previous section. However, in the data mining process, a tuple contained every parameters for each sample in a logged route dataset, is required as input. Thus, the data integration process firstly built up a matrix that converted from $\boldsymbol{R}_k$, then performed a transpose operation to get matrix $\boldsymbol{P}_k$ for completing the data integration process. Each row $r_{k,i}$ in $\boldsymbol{P}_k$ represents a tuple with respect to a sample.

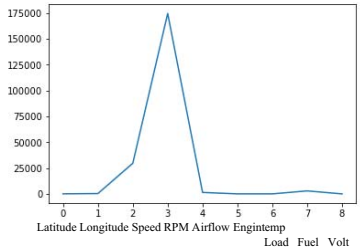
### 3.3 Data selection

The purpose of data selection process is to rank the most useful parameters from $\boldsymbol{P}_k$. It intended to lower the dimension of tuple to reduce the complexity of the data mining process. The methods of Chi-squared( $\chi$ 2) test and Pearson's correlation coefficient are adopted by the data selection process from Scikit-learn package. Fig 4(a) was the Chi-squared ( $\chi$ 2) test result. It ranked engine RPM, velocity, and fuel consumption rate that were the three most useful parameters in order. Fig 4(b) was the Pearson's correlation coefficient result. It ranked velocity, engine RPM, and fuel consumption rate that were the three most useful parameters in order.

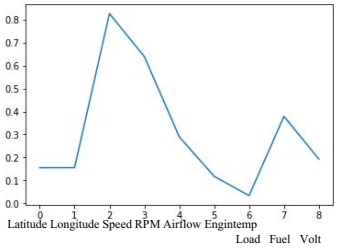
### 3.4 Data transformation

In many related researches of driving behavior analysis mentioned that parameters of acceleration and deceleration, frequency of furious acceleration and deceleration, and number of speeding are also needed. Owing to these parameters are not directly collected and logged in route dataset. They are transformed by specific physical model formulates in data transformation processing. In fact, these parameters are depended on acceleration parameter which can be derived from differential of velocity parameter. Currently, the furious acceleration is defined acceleration greater than 2.74 m/s$^2$ (or 9.864 km/hr/s), and furious deceleration is defined deceleration less than –2.74 m/s$^2$ (or –9.864 km/hr/s). The frequencies of

furious acceleration or deceleration are further counted by the defined thresholds. Then, the acceleration parameter and frequencies of furious acceleration or deceleration are appended into tuple of $\boldsymbol{P}_k$ to create $\boldsymbol{Q}_k$ as the output of data transformation process.




(a) Chi-squared ($\chi 2$) test method




(b) Pearson's correlation coefficient method

Fig. 5 Data selection

3.5 Data mining and knowledge evaluation/display

In the data mining process, taking the velocity and acceleration in $\boldsymbol{Q}_k$ as the input parameters, the decision tree method with statistical kurtosis and skewness are used to analyze the driver behavior. The analyzed qualitative driver behavior classes are defined semantic labels in knowledge evaluation/display process. The kurtosis measures the tailedness of a pdf (probability density function) distribution, which is a fourth standardized moment defined by Eq. 1.

$$k[\boldsymbol{X}] = E\left[\left(\frac{\boldsymbol{X}-\mu}{\sigma}\right)^4\right] = \frac{E[(\boldsymbol{X}-\mu)^4]}{(E[(\boldsymbol{X}-\mu)^2])^2} \quad (1)$$

Where $\boldsymbol{X}$ is a random parameter vector, $\mu$ is the mean, and $\boldsymbol{\sigma}$ is the variance.

A kurtosis is a value without unit. It is also called a kurtosis coefficient. Referring to Fig. 6, a kurtosis coefficient equal to 0 is defined as a normal kurtic, a less than 0 is called a platy kurtic, and a greater than 0 is called a kurtosis letpo kurtic. In reference to velocity-based driving behavior analysis, can be used to distinguish between highway driving (letpo kurtic) or urban driving (platy kurtic)

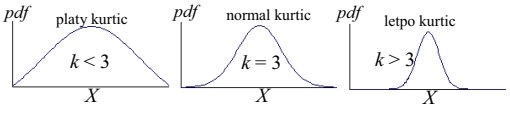



Fig. 6 Classes of kurtosis coefficient

The skewness measures the mean-centric asymmetric distribution of the probability density function which is a third standardized moment defined by Eq. 2.

$$s[\boldsymbol{X}] = E\left[\left(\frac{\boldsymbol{X}-\mu}{\sigma}\right)^3\right] = \frac{E[(\boldsymbol{X}-\mu)^3]}{(E[(\boldsymbol{X}-\mu)^2])^{3/2}} \quad (2)$$

Where $\boldsymbol{X}$ is a random parameter vector, $\mu$ is the mean, and $\boldsymbol{\sigma}$ is the variance.

The skewness is also the unitless value called the skewness coefficient. Referring to the categories illustrated in Fig. 7. A skewness coefficient equal to 0 indicates a symmetric distribution of the probability density function centered on the mean. A skewness coefficient less than 0 is called a negative skewness which probability density function tilted to the right. A skewness coefficient greater than 0 is called positive skewness which probability density function tilted to the left. In reference to acceleration-based driving behavior analysis, the counts of acceleration and deceleration are more frequently when traffic jam. The probability density function tilted to the right.

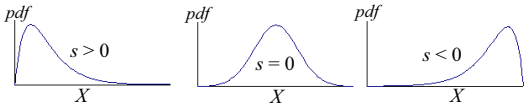



Fig. 7 Categories of skewness coefficient

There are 100 logged route datasets of metropolitan bus are selected to as inputs of data mining process. The decision tree classifier in scikit-learn package has been adopted to realize the function of data mining process based on the kurtosis and skewness of velocity as well as acceleration. Fig. 8 is a trained decision tree by 60 kurtosis coefficients of velocity and acceleration from 100 logged dataset. The left subtree from root represents lower velocity. The left most leaf represents the lower velocity and acceleration which is labeled defensive. On the other hand, the right most leaf represents the higher velocity and acceleration which is labeled aggressive. Referring to Fig. 8, other two leaves are labeled weak defensive and aggressive, respectively. After that, the remainder 40 kurtosis coefficients are used to evaluate the error rate of trained decision tree. It is about 5.8%. Fig. 9 is a trained decision tree by 60 skewness coefficients of velocity and acceleration from 100 logged dataset. Its leaves are labeled defensive, weak defensive, weak aggressive, and aggressive from right to left. The evaluated error rate is about 2.8%.

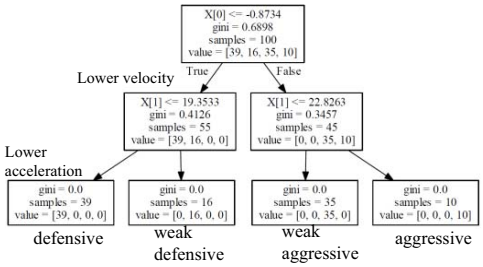



Fig. 8 Trained decision tree based on kurtosis coefficients of velocity and acceleration

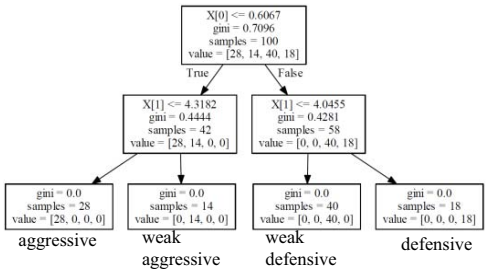



Fig. 9 Trained decision tree based on skewness coefficients of velocity and acceleration

## IV. Conclusion

The analysis of driver behavior has been accomplished through the iterative processes of data cleaning, data integration, data selection, data transformation, data mining, and knowledge evaluation/display based on the datasets of cloud computing platform. Moreover, the iterative processes have been implemented by Python based on Scikit-learn class package as well as the datasets have recorded the dynamical vehicle parameters of every routes from OBD port including velocity, engine RPM, intake air flow rate, engine coolant temperature, engine load, fuel consumption rate, and battery voltage. The route dataset is downloaded from cloud computing platform accordance with the authorized user key and route key. A Kalman filter and redundant zero remover have been realized to remove noisy and redundant data in the downloaded route dataset in data clean process to avoid interference with driver behavior analysis. Data integration process binds the individual vehicle parameter arrays into a matrix for the successive steps of driver behavior analysis.

Then, Data selection process uses the chi-squared ($\chi 2$) test and Pearson's correlation coefficient methods to rank the vehicle parameters in the matrix that designed to reduce the complexity of data mining. The results of data selection process reflects the velocity, engine RPM, and fuel consumption rate are the highest ranks, respectively. Acceleration is another important data for driver behavior analysis which are transformed from the differential of vehicle velocity in data transformation process. Furthermore, statistical kurtosis and skewness of velocity and acceleration are also derived in data transformation process that prepared for data mining process.

In this paper, the decision tree classifier of Scikit-learn have been used for data mining based on the kurtosis and skewness of velocity and acceleration from metropolitan bus route datasets. Significantly, there are four categories being preset as the results of decision tree classifier. Finally, the knowledge evaluation/display process mainly summarizes the classification rules from the completed decision tree classifier, as well as interpret driver behavior category semantics to defensive, weak defensive, weak aggressive, and aggressive to complete the overall processes of driver behavior analysis. The research will continue to involve more parameters to improve the driving behavior classification models that hopes to further optimize the performance of the driving behavior model to enhance the classification of driving behavior degree.

## Acknowledgment

This study is conducted under the “Project for 4G Advanced Business and video services platform” of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

## References

[1] Hyun-Jeong, Y., Shin-Kyung, L., Oh-Cheon, K., "Vehicle-generated data exchange protocol for remote OBD inspection and maintenance," 6th International Computer Sciences and Convergence Information Technology (ICCIT 2011), 81-84, 2011.

[2] Chipan Hwang, Mu-Song Chen, and Hsuan-Fu Wang, "Development of Car Cloud Sensor Network Application Platform," *30th IEEE International Conference on Advanced Information Networking and Applications* (*AINA-2016*), pp. 975-978, Mar. 23-25, 2016, Switzerland

[3] Scikit-Learn, http://scikit-learn.org/stable/, July 2017.

[4] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques, 3e*, Morgan Kaufmann, 2011.