Peer group recognition based on vehicle operation and behavior.

Supervised and unsupervised approach towards peer group recognition and feature space exploration.

Karthik Bangalore Girijeswara

CAISR

Halmstad University, December 4, 2017
Karthik Bangalore Girijeswara: Peer group recognition based on vehicle operation and behavior. Supervised and unsupervised approach towards peer group recognition and feature space exploration. © April 2017
The woods are lovely, dark and deep.
But I have promises to keep.
And miles to go before I sleep.
And miles to go before I sleep.

— Robert Frost
Behavior recognition provides an interesting perspective for understanding the different modes of a system and the influence of each mode under varying conditions. In most of the systems, prior knowledge of different expected behavior is available. Whereas, in an automotive domain, a fleet of vehicle with many external factors influencing each vehicle and an asynchronous performance of each vehicle on road, creates the complexity on analyzing and predicting the exact time segments of vehicles in a fleet exhibiting similar behavior. This thesis focuses on recognizing time segments of vehicles that exhibit similar behavior based on supervised and unsupervised approaches. In supervised approach, classifiers are trained to predict two distinctive operations (highway and in-city). In unsupervised approach, feature space is explored for identification of consistent features and existence of other operations. An unsupervised approach to recognize peer cluster groups is combined with supervised classification results to achieve lower computational complexity.
ACKNOWLEDGEMENTS

I would like to thank my supervisor Slawomir Nowaczyk, Yuantao Fan and Mohamed-Rafik Bouguelia for their meticulous support, exhaustive suggestions and thorough guidance. A big thank you to all those involved directly and indirectly. Last but not the least to Halmstad university and CAISR who made it possible for this work to reach fruition.
# CONTENTS

1 **INTRODUCTION.** 1  
1.1 Data 2  
1.2 Motivation 3  
1.3 Research goals 4  
1.4 Challenges 5  
2 **BACKGROUND.** 7  
3 **METHODOLOGY.** 11  
3.1 Supervised classification approach 11  
3.1.1 Preprocessing stage 13  
3.1.2 Classifiers 18  
3.1.3 Validation 18  
3.2 Unsupervised approach 19  
3.2.1 Feature space exploration. 19  
3.2.2 peer group recognition. 27  
4 **DISCUSSION OF RESULTS.** 35  
4.1 Feature space exploration for vehicle dataset 36  
4.2 Test Case 36  
5 **CONCLUSION** 43  
5.1 future work 43  

**BIBLIOGRAPHY** 45
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Different multivariate time series data</td>
<td>3</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Histogram plots for highway and incity operations.</td>
<td>4</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Flow chart of supervised approach.</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4</td>
<td>A small section of encoded gps longitude and corresponding timestamp values.</td>
<td>12</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Map plot of multiple routes from GPS locations gathered from the bus. Routes with pink shade correspond to highway operation and routes with green shade are the in-city operations.</td>
<td>13</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Vehicle speed time series data recorded for a duration of approximately 15 minutes</td>
<td>14</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Vehicle speed time series data recorded for a duration of approximately 15 minutes with corresponding data segments</td>
<td>15</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Top down algorithm flow chart for segmentation of time series data.</td>
<td>15</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Data segmentation of vehicle speed.</td>
<td>16</td>
</tr>
<tr>
<td>Figure 10</td>
<td>ROC and AUC for different classifiers based on complete feature set</td>
<td>20</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Flow chart of feature space exploration.</td>
<td>21</td>
</tr>
<tr>
<td>Figure 12</td>
<td>t-SNE based visualization of different cluster similarity over combinations of clustering algorithms and features.</td>
<td>22</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Gaussian fit for identifying clusters for each clustering algorithm results shown in Figure 12</td>
<td>23</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Comparison of clustering algorithm performance for different feature set</td>
<td>24</td>
</tr>
<tr>
<td>Figure 15</td>
<td>ROC and AUC for different classifiers based on the feature set proposed from the unsupervised approach.</td>
<td>25</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Extra-trees classifier based feature importance (supervised approach).</td>
<td>26</td>
</tr>
<tr>
<td>Figure 17</td>
<td>PCA reduced v-matrix for synthetic dataset.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Silhouette index computed for different features in v-matrix</td>
<td>28</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Flow chart of affinity propagation and Silhouette index based peer group generation.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Affinity Propagation clustering results for different buses.</td>
<td>30</td>
</tr>
</tbody>
</table>
Figure 21  Output of peer clusters with similar behavior based on SI metric. 31
Figure 22  Percentile based outliers detection on SI metrics 32
Figure 23  Peer cluster suggestion based on CI ratios. 33
Figure 24  PCA reduced v-matrix for vehicle dataset. 36
Figure 25  v-matrix based feature evaluation on vehicle dataset 37
Figure 26  Affinity propagation clustering of bus 369 38
Figure 27  Affinity propagation clustering of bus 372 39
Figure 28  Affinity propagation and Silhouette index based peer group recognition. 40
Figure 29  Percentile based outlier detection of Silhouette index metrics. 41
Figure 30  Proposed clusters based on class ratio metric. 42
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Different Classifier performance evaluation on the initial selected feature set.</td>
<td>19</td>
</tr>
<tr>
<td>Table 2</td>
<td>Performance evaluation of different clustering algorithms on feature sets corresponding to both most similar clustering result and the outlying clustering result.</td>
<td>22</td>
</tr>
<tr>
<td>Table 3</td>
<td>Evaluation of different Classifier performance based on the feature set proposed from feature space exploration technique.</td>
<td>25</td>
</tr>
<tr>
<td>ACRONYMS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
<td></td>
</tr>
<tr>
<td>CAISR</td>
<td>Center for Applied Intelligent Systems Research</td>
<td></td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
<td></td>
</tr>
<tr>
<td>SI</td>
<td>Silhouette Index</td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>Class Index ratio</td>
<td></td>
</tr>
<tr>
<td>VACT</td>
<td>Volvo Analysis and Communication Tool</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
<td></td>
</tr>
<tr>
<td>TSNE</td>
<td>t-distributed stochastic neighbor embedding</td>
<td></td>
</tr>
<tr>
<td>ARI</td>
<td>Adjusted Rand index</td>
<td></td>
</tr>
<tr>
<td>PLA</td>
<td>Piecewise Linear Approximation</td>
<td></td>
</tr>
<tr>
<td>Q₁</td>
<td>First Quartile</td>
<td></td>
</tr>
<tr>
<td>Q₃</td>
<td>Third Quartile</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector machine</td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>K Nearest Neighbor</td>
<td></td>
</tr>
<tr>
<td>GPC</td>
<td>Gaussian Process Classifier</td>
<td></td>
</tr>
<tr>
<td>SGD</td>
<td>Stochastic Gradient Decent</td>
<td></td>
</tr>
<tr>
<td>GNB</td>
<td>Gaussian Naive Bayes</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-Layer Perceptron</td>
<td></td>
</tr>
</tbody>
</table>
INTRODUCTION.

Machine learning and data mining techniques provide better insight on patterns, or reveal some interestingness present in the data. This new found information via data mining can be used in many applications. As time passes by, we encounter more accumulated data which are not analyzed, that may contain important and useful information. In the field of automotive industry, evolving technology in the market and introduction of new functions and features in the vehicle, lead to integration of more physical systems in the vehicle, with computer-based systems. Analyzing the data gathered from such integrated systems, to recognize peer groups of time segments based on vehicle behavior and operation performed is proposed in this thesis.

Predictive maintenance and fault detection is a solution enforced in evolving automobile industry that would avoid unplanned stops. Fault detection would be much more effective if each vehicle in a fleet is compared to other vehicles performing similar operations. A vehicle operation in this thesis refers to different tasks that a vehicle performs when on road. Examples of different vehicle operations are workshop visits, distinguishing between highway and in-city operation, or finding uphill and downhill driving as well as in more abstract description such as, is the bus in regular line traffic or performing some other duty, different conditions the bus is in such as, high density traffic, in snow covered road, driving in rain, etc. The main focus on classifying different vehicle operation is to have a similar performance in subsystems of different vehicles in a fleet performing same operations. The reason to why we don’t completely rely on this operation classification(supervised approach) is mainly due to multiple external factors that influence the performance of a vehicle. External factors such as driver performance, different models of buses in a fleet have varying performance, a vehicle that is serviced recently has a variation in performance in comparison to another vehicle exhibiting the same behavior. Hence, unsupervised approach is also explored in generating peer groups of time segments from vehicles with similar behavior in a data driven approach irrespective of the operation being performed.

Vehicle operation classification is used to characterize the difference in behavior of the vehicle. A supervised approach of recognizing different behavior is by classifying the operation of vehicle. Two such important and distinct operations are highway and in-city.
broader perspective towards analyzing the data for identification of different behavior that exist in the time series data is by unsupervised approach. Unsupervised approach examines the feature space for existence of other operations apart from highway and in-city. Feature space exploration is also used for identifying consistent feature set that improves the performance of both supervised classifiers and clustering algorithms. In this thesis, we explore expert knowledge based(supervised) and data driven based (unsupervised) approaches for identifying time segments of the vehicles performing similar operation or exhibiting similar behavior(Peer group). peer group in the context of this thesis refers to objects(time segments) that exhibit similar performance in the vehicle.

1.1 DATA

Data used in this thesis was accumulated from a fleet of 19 buses operating on the west coast of Sweden, near the city Kungsbacka. Data collected from the bus spans over a duration of 3 years. Each vehicle in the fleet had the ability to record and transmit data stream from on board sensors, control signals and commands, with the help of VACT (Volvo Analysis and Communication Tool). VACT uses telematics technology to remotely transmit data to a back office server. This setup was installed so that each bus in the fleet had the ability to transmit compressed representation of on-board data, that can be analyzed to predict faults and potential failures of vehicle subsystems. Data accumulated from each bus consists of approximately 100 time-series signals, e.g., GPS, Vehicle and engine speed, different pedal positions, etc. Each of the signals were logged at a frequency of 1 Hz.

Knowledge discovery or Data Mining can be defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [11]. Time-series data is one such data that contains a sequence of values obtained at different time instances. A sequence composed by a series of nominal symbols from a particular alphabet is usually called a temporal sequence and a sequence of continuous, real-valued elements, is known as a time series data [2].

Determining the behavior of a variable over time, provides a great insight in predicting the outcome of a system. Time-series data is present everywhere. Most of the data from the real world are time series data, data that is obtained periodically at different time instances. Time-series data is a very important factor in Weather prediction, speech recognition, financial engineering like sales prediction and stock market analysis, economic forecasting, astronomy, hydrology, health and many more domains. Tasks related to time-series are segmentation[17], prediction[5], classification[27], clustering[21],
1.2 Motivation

The main motivation for classification of vehicle operation and behavior based peer grouping, is its use in fault detection of predictive maintenance. In automotive industry, unplanned stops and breakdowns of vehicles is very problematic. To avoid such breakdowns of vehicle due to faults that may arise in the vehicle, a predictive maintenance system would monitor the data gathered from on-board sensors. The data from on-board the vehicle is gathered using a tool called VACT (Volvo analysis and communication tool). VACT tool provides the vehicle, ability to transmit the data available on board acquired from various sensors and control units.

In previous research,[10] a predictive fault detection of air compressor in bus fleet data using histograms was proposed. In this technique,
the histogram data of the wet tank air pressure over a period of 1 week of each bus, was compared with the average histogram data of the remaining fleet of buses. The distance difference of histograms between a bus and the remaining fleet would represent the anomaly metric. Although the result of such anomaly metric did represent changes when air compressor of a bus was replaced or serviced, it also provided false alarms or high anomaly measure even when the compressor was not faulty. One of the reasons behind this error in prediction would be the difference in behavior of a certain system during different operations of the vehicle. The performance of a vehicle varies when on highway and during in-city operation as shown in the histogram Figure 2. Figure 2 clearly depicts 2 different histograms (different probability distribution) of the air compressor data during 2 different operations. Hence comparison of a vehicle with more in-city operation over the week with vehicle having more highway operation (or vice-versa) will yield an anomalous behavior even when there is no fault in the vehicle.

Thus, it is important that detecting anomalies in a fleet is subjective to comparing the vehicles performing same operation or exhibiting similar behavior. Since labels to some important operations don’t exist, unsupervised approach is also necessary to analyze for important features and peer grouping.

![Histogram plots for highway and incity operations](image)

Figure 2: Histogram plots for highway and incity operations

1.3 RESEARCH GOALS

At any time instance in a fleet of buses, there exists multiple buses performing different operations, for varying duration of time. To compare buses with similar behavior, for better predictive maintenance, peer groups of adaptive time segments should be generated to group segments with same operations or similar behavior.

The research goals of this thesis are:
• Train and validate classifiers for accurate vehicle operation classification (Supervised approach).

• Unsupervised approach for feature space exploration.

• Explore data driven approach in recognizing peer groups of time segments based on vehicle behavior (Unsupervised approach).

• Explore the database for existence of operations other than just highway and in-city.

• Propose a methodology that is efficient in recognizing peer groups for the large data set.

To achieve the above mentioned goals, there are a few challenges that we should overcome as discussed in the next section.

1.4 Challenges

Challenges that were faced during the process of peer group recognition are:

• Huge amount of data: At an average, each bus is on road for 5 hours every day. Since the data is gathered at a frequency of 1 hertz, we end up with 18,000 time-series data from each signal of every bus. To overcome this challenge, data segmentation algorithm is used for reduction of data as explained in section (3.1.1) providing lower computational complexity.

• Asynchronous time series data: each signal from on-board the bus is gathered at different time periods. Data segmentation provides a window of time series data that has linear change in time. Hence, comparison of different time signals with respect to the window of time from the data segments, eases the comparison of asynchronous signals.

• In-adequate operation labels: There are many operations that a vehicle can perform. In most situations a vehicle might be performing 2 or more operations simultaneously, one example of such scenario is a vehicle on an highway and going uphill in dense traffic. Ground truth (labeled data) to train and validate classifiers on all such operations is not available in the data set. Hence we also implement unsupervised approach to explore the dataset to find clusters that determine the time segments of vehicles during which the behavior are similar irrespective of the operation label being unknown.

• High computational complexity: generation of peer groups from multiple clusters of each bus to the rest in a fleet is a tedious and
high computational task. Hence in this thesis a more efficient way of generating such peer groups are proposed by combining the supervised classifier results along with unsupervised approach.
BACKGROUND.

The main type of data focused in this thesis is time-series data. The database was gathered from a fleet of buses, each bus transmitting hundreds of signal at a frequency of one second. Dealing with such huge amount of data requires data reduction and dimensionality reduction techniques for efficient implementation of different techniques and lower computational complexity. As per the article, Effect of segmentation on financial time series pattern matching \[33\], time series data segmentation reduce the dimensionality of the time series while preserving essential features and characteristics of the original time series.

Data segmentation of time series data reduces memory space consumption. With low maximum error on linear piecewise approximation \[20\] Algorithmic methods for segmentation of time series is an article that explains different time series data segmentation that exist and evaluate the performance of each algorithm. Top down algorithm had the better results compared to bottom up and sliding window algorithms.

High dimensionality reduction is an important technique in understanding the structure of the data and to lower computational complexity. Two most widely implemented dimension reduction techniques are Principal component analysis (PCA)\[31\] and t-distributed stochastic neighbor embedding (TSNE) \[1\]. Both the techniques has its advantages and disadvantages. PCA is can only capture linear structures in high dimension features, whereas TSNE captures the non-linear structure present in high dimension space. However, TSNE is non parametric learning algorithm, which means it does not learn a function to reduce new high dimensional data to the low dimensional space. Also, in PCA the optimal solution to the problem is guaranteed. whereas, TSNE has a non-convex objective function that uses gradient descent in optimization leading to a different solution at each execution \[32\]. Hence, considering the the above conditions, PCA is implemented in the task where testing and validation of new data is necessary such as clustering and classification. TSNE technique is used for the task of feature space exploration to validate the existence of other significant operations and consistent feature recognition.

Time series data is a very commonly observable data type in many fields of application. One such important domain that deals with such
multivariate time series data is stock market. [18] is a research in this domain with an objective most similar to goal of this thesis. [18] is a stock fraud detection using peer group analysis. The key ingredient for fraud detection is peer group analysis of similar objects. Peer groups are objects that have similar pattern of change over time. Their technique involves 2 stages: 1) identifying and building peer groups, and 2) detecting anomalous behavior in the constructed peer groups. Peer groups were generated based on similarity between time series.

In this thesis the objective of peer grouping objects with similar behavior is similar, but two time series signal from different buses having similar behavior does not continue to have the same change in pattern since each bus in the fleet is not subjected to same experience. Also, there is no prior knowledge which of the multivariate time series signals of a bus would be used to justify the similarity in behavior.

In article [10] a histogram based fault detection system is proposed. In this technique, the histogram data of the wet tank air pressure over a period of 1 week of each bus, was compared with the average histogram data of the remaining fleet of buses. The distance difference of histograms between a bus and the remaining fleet would represent the anomaly metric. Although the result of such anomaly metric did represent changes when air compressor of a bus was replaced or serviced, it also provided false alarms or high anomaly measure even when the compressor was not faulty. To reduce the errors in such anomaly detection is a motivation for this thesis.

Sung-Hyuk in his article [4], Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, provides a summary of various similarity measures that are applicable for comparison of 2 probability density function. There are various distance measures that exist in calculating the similarity between probability density functions. Sung in his article provides 4 prominent distance measures, which are, Euclidean and Chebyshev distances from Minkowski family, cosine distance from inner product family and Hellinger distance from Fidelity family. This computation of similarity measure is used to indicate that there is a huge difference in performance of a vehicle depending on the operation (Highway and In-city).

Evaluation of a system is an important step in recognizing the performance and also in validation or comparison of different techniques. There are many evaluation metrics used in supervised and unsupervised techniques [23]. In supervised classifier, metrics like the following are used for evaluating the performance of a classifier [24]:

- Precision : This metric is the ability of a classifier not to misclassify a negative sample to be positive.
• Recall: This metric indicates the ability of classifying all the positive samples correctly.

• F1 score: This metric is a weighted harmonic mean of precision and recall.

• Receiver operating characteristic (ROC): ROC curve is a commonly used technique for visualization of binary classifier. This visualization graph is generated by plotting the true positive rate against the false positive rate of a classifier at different thresholds. Also, accuracy is measured by the Area under the ROC curve (AUC) [16].

In unsupervised approach, metrics like the following are used for evaluating the performance of a clustering algorithm [25]:

• Adjusted Rand index (ARI): ARI is a metric that measures the similarity in assignment of the clustering labels to the ground truth.

• Homogeneity: This metric indicates if the clustering algorithm result contains clusters of data points belonging to a single class.

• Completeness: This metric indicates if the clusters generated from the clustering algorithm contain all the data points of a given class.

• V-measure: V-measure is a metric similar to mutual information that is computed by generating the harmonic mean between homogeneity and completeness.

• Silhouette index (SI): A Silhouette metric reveals the overlapping of clusters in the data space by using mean intra-cluster distance and nearest cluster distance for each observation in the data space [26].

Extremely Randomized classifiers, also known as extra classifiers is a very common technique for feature selection technique Pedregosa et al. [22]. Extra classifiers are used to fit a number of randomized decision trees over sub-samples of dataset. A tree based estimator for a supervised technique in evaluating feature importance. Recently this technique is also been used to select acoustic features for emotion recognition Cao et al. [3]. This technique is used for verifying the performance of unsupervised feature space exploration in synthetic dataset.

There are many clustering algorithms used in the data driven approach for identifying different cluster densities and each of those algorithms have its limitations [28]. One of the most commonly found limitation in many such algorithms, is to have prior knowledge in the number of clusters that the data space contains. Since in this thesis we
try to explore all possible clusters that exist in the data space, affinity propagation is an interesting algorithm that chooses the number automatically based on the data provided Frey and Dueck [12]. This algorithm creates clusters in the data space by passing messages between pairs of data points until convergence. Each cluster in the data set is represented by an exemplar (most representative data point among a cluster).
Methodology of this thesis includes exploring both supervised and unsupervised techniques as explained in the following sections.

### 3.1 Supervised Classification Approach

Supervised vehicle operation classification is focused on 2 main operations 1) Highway and 2) In-city operations. The choice of analyzing these two operations was because of the availability of ground truth that can be used to generate training and validation data. Also, these operations exhibit a distinctive change in vehicle performance. As observed in related work [10] the behavior of air compressor component within the vehicle depends mainly on the highway/in-city operation of the bus. From expert knowledge, average pressure will be lower in-city than on highway due to frequent braking.

First step towards supervised classification of vehicle operation is to generate labeled data for training and validation purposes. One of the on-board signals from bus contains GPS positions recorded at a frequency of one second and each recording of a signal constitutes a timestamp along with it Figure 4, which can be decoded to get the exact date and time of any recorded signal. Plotting the GPS locations over time provides the route of the bus taken as shown in Figure 5. Approximately 140 minutes of recorded data during each operation is extracted for different buses.

Since each signal from on-board is a time series data, every GPS location plotted on the map has a time stamp to it. Once the labels for GPS timestamps are generated, the signals corresponding to these timestamps are extracted as training data for specific classes. For modeling the behavior of a vehicle, we focus on using control signals and signals that are ineffective to any malfunction of internal components of the vehicle. The signals used for modeling the behavior of the vehicle are:

- Vehicle speed.
- Selected gear.
- Brake pedal position.
- Accelerator pedal position.
- Engine fuel temperature.
Figure 3: Flow chart of supervised approach.

Figure 4: A small section of encoded gps longitude and corresponding timestamp values.
Supervised classification approach

Figure 5: Map plot of multiple routes from GPS locations gathered from the bus. Routes with pink shade correspond to highway operation and routes with green shade are the in-city operations.

- Fuel Rate.
- Engines speed.
- Relative speed of right wheel.
- Relative speed of left wheel.

Once the data are labeled and training data is ready, preprocessing stage begins.

3.1.1 Preprocessing stage

Three important steps in the preprocessing stage of training a classifier are:

- Data segmentation.
- Feature extraction.
- Feature selection.

Preprocessing stage provides refined data to classifiers rather than the raw data, to improve classification accuracy and lower computational complexity. Data segmentation reduces the data space into piecewise linear segments corresponding to a sequence of data points with similar change over time. Thus, reduction of data space lowers
the computation complexity and easy representation of large data. Once the segments are generated, we focus on extracting features of these segments that may help in classification process.

3.1.1.1 Data segmentation

One of the challenges is to represent huge data, as shown in Figure 6 a signal (vehicle speed) recorded for approximately 15 minutes consumes around 870 data points to represent. Usually time series data have very small change in value with respect to time, hence a piecewise linear approximation (PLA) of the original signal would contain a time series on linear segments that represents the original signal Figure 6. The complete signal of 80 data points is now represented with just 66 data segments. As discussed in Section 2 there are multiple algorithms to generate PLA segments.

Top down segmentation algorithm is more suitable for segmenting time series data among various other algorithms. Figure 8 shows the flow chart of top-down algorithm for piecewise linear approximation of time series data. The outcome of segmentation procedure, approximates time series, over non-overlapping segments, presented in the form of straight lines as shown in Figure 9. Top-down algorithm method is to recursively cut the current data segment into 2 sub segments, until some error threshold is reached. The error threshold used here is 1 (very low) for the purpose of fine representation. Also, since the variation of data is high during in-city operation compared to highway operation the number of segments required to represent in-city data is more. Hence in-city data segments are more compared to highway, and average segment duration (length of the segment) is very less for in-city operations.

Figure 6: Vehicle speed time series data recorded for a duration of approximately 15 minutes
3.1 Supervised Classification Approach

Figure 7: Vehicle speed time series data recorded for a duration of approximately 15 minutes with corresponding data segments.

Figure 8: Top down algorithm flow chart for segmentation of time series data.

As observable from Figure 9 the data points between 850 to 910 are values that fluctuates in and around 89.8 to 90.1, that is, approximated...
by a single segment. This is an example of data reduction while still preserving the essential features of the original time series data.

3.1.1.2 Feature extraction

As the data segments contain linear approximation, features of those linear representations are:

- Slope: capture the temporal aspect, that is change in value over time for the given segment duration.
- Segment duration: length of the segment.

We also focus on extracting other statistical features of time series segments that would be further used in creating a model for classification. Other statistical features are:

- Mean: it is the average value of the overall segment
- Standard deviation: metric that quantifies the variation of the data.
- Minimum: minimum value of the segment
- Maximum: maximum value of the segment
- First Quartile: The first quartile (Q1) is defined as the middle value between the minimum and the median of the data segment[35].

Figure 9: Data segmentation of vehicle speed.
• Median: 2nd quartile, middle value between the maximum and minimum of the data segment.

• Third Quartile: is the middle value between median and the maximum value of the data segment.

• Skewness: it is the measure of lack of symmetry in data.

• Kurtosis: metric that conveys whether the data is heavy tailed or light tailed relative to a normal distribution.

Apart from these, gradient feature extraction on all the above-mentioned features would further generate features that convey information about the temporal changes in the data.

Once all 9 statistical features are extracted from segments of 11 different signal and their gradients are computed, feature selection is used to select a few of the 164 features that would be relevant for classification.

3.1.1.3 Feature selection

The focus of this stage is to reduce dimensionality of 164 features collected. L1 based feature selection is used to select a set of features from the initial data set based on the feature weight coefficient. Feature weight is a metric that would approximate the importance attribute of a certain feature based on recursively training an estimator that considers smaller set of features in each iteration. Linear models are used to interpret the features and for the process of selecting a strong subset of features for improving model performance. Hence the different signals and features selected for the requirement of classifying highway and in-city are as mentioned below:

1. Selected Signals are:
   • Vehicle speed.
   • Engine speed.
   • Accelerator pedal position.
   • Engine Fuel temperature.
   • Fuel rate.
   • Segment duration.

2. Different features of the above mentioned signals are:
   • Minimum.
   • Maximum.
   • Gradient.
   • First quartile.
- Median.
- Third quartile.

3.1.2 Classifiers

Once the training and validation data is ready after feature selection, standard machine learning classifiers are trained for prediction and validation. The classifiers explored are:

- Support Vector machine (SVM), is a supervised method used widely for its effective performance on data with high dimensional spaces (even in cases where the number of observations are less than the number of dimensions [29]).

- K nearest neighbor (KNN), classification technique is an instance based type of learning, k value is responsible for the distinction between boundaries and a large value of k, suppresses noise[6].

- Stochastic gradient decent is a an efficient approach towards classifying large scale observations dataset[30].

- Gaussian Process Classifier (GPC), provides a gaussian process for a latent function to obtain probabilistic based classification.

- Gaussian Naive Bayes (GNB) estimates the parameters for the features based on maximum likelihood, with an assumption that it is Gaussian.

- Decision tree (DT) based classifier, is a predictive model generated based on the information gain available for each feature in the training dataset.

- Multi-Layer Perceptron (MLP) classification trains the neural network using Backpropagation technique. The model optimizes the loss function using stochastic gradient descent[14].

- Random forest classifiers is based on randomized decision trees, where each tree in the ensemble creates a tree by splits chosen on a random subset of features[7].

To generalize the model, data from multiple buses are extracted during different periods of the year. A dataset of 22 features in each 3160 samples were generated from 4 different buses with multiple routes throughout a year is used to train and validate the models. Validation of these models are based on Accuracy, ROC and Precision.

3.1.3 Validation

Validation is an important stage in analyzing the performance of the models built. Cross validation is technique that validates each model
based on a separate testing data that the model was not trained on. Thus, the true accuracy of a model is recognized based on unseen data by the model.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Cross validation accuracy</th>
<th>Operation (Class)</th>
<th>Precision</th>
<th>Recall</th>
<th>f1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>88.433</td>
<td>Highway</td>
<td>0.71</td>
<td>0.66</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>KNN</td>
<td>93.715</td>
<td>Highway</td>
<td>0.90</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.94</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>SGD</td>
<td>85.61</td>
<td>Highway</td>
<td>0.72</td>
<td>0.39</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.87</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>GPC</td>
<td>93.715</td>
<td>Highway</td>
<td>0.91</td>
<td>0.74</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.94</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>GN Bayes</td>
<td>78.688</td>
<td>Highway</td>
<td>0.47</td>
<td>0.88</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.96</td>
<td>0.77</td>
<td>0.85</td>
</tr>
<tr>
<td>Decision tree</td>
<td>93.897</td>
<td>Highway</td>
<td>0.85</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>MLP</td>
<td>93.715</td>
<td>Highway</td>
<td>0.81</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.97</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>Random Forest</td>
<td>95.355</td>
<td>Highway</td>
<td>0.89</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Evaluation of supervised classifiers considering complete feature set

Table 1: Different Classifier performance evaluation on the initial selected feature set.

3.2 UNSUPERVISED APPROACH

Unsupervised approach is the data driven method of analyzing and clustering of data. Unsupervised approach focuses on exploring different feature space and recognizing peer cluster groups with similar behavior.

3.2.1 Feature space exploration.

This step in unsupervised approach is not only crucial for identifying different useful consistent features but also to explore for existence of other vehicle operations. Feature space exploration is implemented by analyzing clustering results of different algorithms over different combinations of feature. Flow chart of this process is shown in Figure 11. In this stage, different combination of clustering results are clustered again to identify interesting features and clustering algorithms.

The first step here is to generate different clustering results over 7 different features. The different clustering algorithms used here are:
K means clustering.

- Gaussian Mixture Model clustering.
- Spectral Clustering.

The combination of 7 different signals yield 120 different feature space, and implementation of three different algorithms as mentioned above gives 320 different clustering results.

Once these 320 cluster sets corresponding to clustering result from different combination of algorithms and feature space are generated, we then evaluate the similarity of each clustering result with the rest. To compute this similarity between different cluster sets, V-measure metric is used. V-measure metric is the harmonic mean between completeness and homogeneity. So this V-measure computation of each clustering result with the rest gives a $320 \times 320$ matrix, where each row of the matrix represents the clustering result similarity between one and rest. This matrix of high-dimensional data that describes the similarity of different clustering result is referred to as v-matrix.

For the purpose of reduction and visualization of this high-dimensional V-measure matrix, t-Distributed Stochastic Neighbor Embedding (t-
Figure 11: Flow chart of feature space exploration.

SNE) technique is implemented Figure 12. As shown in Figure 12 there are a few different clusters for each algorithm and also multiple clusters within each algorithm. This shows that there is a difference in performance of each algorithm with respect to the selected feature space and there exists other operations that certain features explore.

Once we can visually identify the number of clusters that exist for each algorithm, Gaussian mixture model is used to group data points within high density regions as shown in figure.

By computing the centers of those clusters with high density and backtracking to the original database before t-SNE dimensionality reduction, most consistent features of corresponding algorithms are recognized. To prove this, the features of data point with best fit for the high density clusters corresponding to each clustering algorithm in Figure 13 is considered as a good features set. Whereas, clustering results with a bad fit to the same gaussian is considered to have a feature space with worse impact on performance. Figure 14 and Ta-
Methodology.

Figure 12: t-SNE based visualization of different cluster similarity over combinations of clustering algorithms and features.

Table 2 is a process of evaluation to prove that the clustering results that are most similar (center of gaussians in Figure 13) for all the combinations has a better and consistent performance in comparison to the most outlying clustering results.

<table>
<thead>
<tr>
<th>Clustering algorithm</th>
<th>Homogeneity</th>
<th>Completeness</th>
<th>V-measure</th>
<th>ARI</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral clustering*</td>
<td>0.03</td>
<td>0.021</td>
<td>0.0250</td>
<td>-0.042</td>
<td>-0.03</td>
</tr>
<tr>
<td>Spectral clustering**</td>
<td>0.707</td>
<td>0.780</td>
<td>0.742</td>
<td>0.853</td>
<td>0.309</td>
</tr>
<tr>
<td>K-means clustering*</td>
<td>0.105</td>
<td>0.087</td>
<td>0.095</td>
<td>-0.101</td>
<td>0.314</td>
</tr>
<tr>
<td>K-means clustering**</td>
<td>0.311</td>
<td>0.566</td>
<td>0.398</td>
<td>0.481</td>
<td>0.714</td>
</tr>
<tr>
<td>Gaussian Mixture Model*</td>
<td>0.437</td>
<td>0.605</td>
<td>0.507</td>
<td>0.633</td>
<td>0.568</td>
</tr>
<tr>
<td>Gaussian Mixture Model**</td>
<td>0.963</td>
<td>0.972</td>
<td>0.967</td>
<td>0.988</td>
<td>0.639</td>
</tr>
</tbody>
</table>

*performance on feature set having the least similarity.
**performance on feature set corresponding to the best fit in similarity.

Table 2: Performance evaluation of different clustering algorithms on feature sets corresponding to both most similar clustering result and the outlying clustering result.
Figure 13: Gaussian fit for identifying clusters for each clustering algorithm results shown in Figure 12

To evaluate the change in performance of each clustering algorithm with different feature space, metrics like silhouette index (SI), homogeneity, completeness, Adjusted Rand Index (ARI) and V-measure are used (Table 2). Apart from the influence of different feature set in clustering algorithms, the feature set with most similar clustering result was used in training and validating the supervised classifiers. As shown in Figure 15 and Table 3 in comparison to the initial classifier evaluation (Figure 10 and Table 1) considering all the features based on L1-based feature selection technique, the feature set proposed from this unsupervised approach yields better performance for all the classifiers.

V-measure matrix (v-matrix) of different clustering results can be further used to identify features that are important by analyzing the PCA-dimensionality reduction of high dimensional v-matrix with respect to individual features.
3.2.1 Methodology.

(a) Spectral clustering on the feature set having the best fit in similarity among other combinations

(b) Spectral clustering on the feature set having the least similarity to other clustering results

(c) K-means clustering on the feature set having the best fit in similarity among other combinations

(d) Spectral clustering on the feature set having the least similarity to other clustering results

(e) Gaussian mixture model on the feature set having the best fit in similarity among other combinations

(f) Gaussian mixture model on the feature set having the least similarity to other clustering results

Figure 14: comparison of clustering algorithm performance for different feature set

3.2.1.1 V-matrix based feature space evaluation on synthetic dataset

To verify this unsupervised approach of evaluating features, we examine this technique on a synthetic dataset generated for ground-
Table 3: Evaluation of different Classifier performance based on the feature set proposed from feature space exploration technique.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Cross validation accuracy</th>
<th>Operation (Class)</th>
<th>Precision</th>
<th>Recall</th>
<th>f1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>90.981</td>
<td>Highway</td>
<td>0.78</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.93</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>KNN</td>
<td>95.253</td>
<td>Highway</td>
<td>0.89</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>SGD</td>
<td>90.506</td>
<td>Highway</td>
<td>0.81</td>
<td>0.62</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.92</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>GPC</td>
<td>94.77</td>
<td>Highway</td>
<td>0.90</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>GN Bayes</td>
<td>86.787</td>
<td>Highway</td>
<td>0.60</td>
<td>0.85</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.96</td>
<td>0.87</td>
<td>0.92</td>
</tr>
<tr>
<td>Decision tree</td>
<td>96.1234</td>
<td>Highway</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>MLP</td>
<td>95.094</td>
<td>Highway</td>
<td>0.92</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Random Forest</td>
<td>96.439</td>
<td>Highway</td>
<td>0.93</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-city</td>
<td>0.97</td>
<td>0.99</td>
<td>0.98</td>
</tr>
</tbody>
</table>

performance of classifiers for the consistent features

Figure 15: ROC and AUC for different classifiers based on the feature set proposed from the unsupervised approach.
truth. The synthetic dataset consists of 7 features (0-6), of which 5 are informative and 2 are non-informative. The dataset is based constructed based on 4 clusters. Each cluster is located in the hypercube subspace of respective informative features. To include covariance in the dataset, each feature is combined linearly within each cluster. Figure 16 is the feature importance evaluation based on extra-trees classifier, a supervised approach used as ground truth. As observable in Figure 16 the feature 1 is most informative among all. Each combination of feature results in different clustering results and this is captured in the v-matrix. Once the v-matrix is computed for the synthetic dataset, PCA dimensionality reduction is used for visualization as shown in Figure 17. Each data point in Figure 17 represents a clustering result for unique combination of feature space.

Figure 16: Extra-trees classifier based feature importance (supervised approach).

To evaluate each feature importance, the PCA data space in Figure 17 is explored repeatedly. For each feature evaluation, silhouette index is computed between the clustering results containing a particular feature against the clustering results that do not (Figure 18). Silhouette index provides a relative metric of feature importance in PCA reduced v-matrix space. As expected from the ground truth (Figure 16), feature 1 has the highest silhouette metric and features 3,
3.2 UNSUPERVISED APPROACH

In supervised approach, we discuss about classification of time segments as either an in-city or highway operation, on the contrary given any instance, the vehicle may be performing multiple other operations along with in-city and highway. For example, a vehicle on highway can also be in traffic and going uphill at the same time. As seen in Figure 12 there are multiple clusters within the same clustering technique. Hence there exist certain features that indicate other underlying operations apart from just highway and in-city. Hence in the next section, to explore unknown operations, affinity propagation technique and silhouette index based peer group recognition is discussed.

3.2.2 peer group recognition.

peer group recognition is a data driven approach of identifying clusters without any prior assumption about the number of clusters. This is because each vehicle in a fleet has a different experience when on road and the behavior might be similar to some operations for a certain bus, whereas the same operations may exhibit varying behavior for a different bus in the same fleet. Hence this approach is based on Affinity propagation clustering and silhouette coefficient. The flow
(a) Silhouette index for feature 0 is 0.128

(b) Silhouette index for feature 1 is 0.624

(c) Silhouette index for feature 2 is 0.092

(d) Silhouette index for feature 3 is -0.001

(e) Silhouette index for feature 4 is 0.278

(f) Silhouette index for feature 5 is -0.002

(g) Silhouette index for feature 6 is -0.006

Figure 18: Silhouette index computed for different features in v-matrix

[December 4, 2017 at 7:23 – classicthesis ]
chart of this approach is shown in the Figure 19. The first step of this approach is to reduce dimensionality for data visualization and lower computation complexity as explained in the following sections.

Figure 19: Flow chart of affinity propagation and Silhouette index based peer group recognition.

3.2.2.1 **PCA dimensionality reduction.**

Since the original data set consists of many features, PCA analysis is carried out for reduction of this high dimensionality, while retaining most information like how the original data is distributed in the PCA reduced data set.

3.2.2.2 **Affinity propagation.**

Affinity propagation technique of clustering is not biased to identify predefined number of clusters, this algorithm iteratively passes messages(responsibility and availability) between pairs of samples and arrives at a convergence of the number of clusters based on the data. Since we are not interested in identifying only highway and in-city operations, affinity propagation technique is implemented on PCA reduced data of all the signals in the feature set. Figure 20 shows the
implementation results of affinity propagation clustering over different bus data.

![Affinity Propagation clustering results for different buses.](image)

Figure 20: Affinity Propagation clustering results for different buses.

This clustering algorithm creates different amount of clusters for different buses, since there may be similar behavior in a bus for multiple operations leading to small number of clusters. whereas, in some scenarios like bus 371 of Figure 20 there maybe large number of clusters due to a significant change in performance of the bus for different operations. This change in behavior or similar behavior for different operations in vehicles are mainly due to two reasons:

- Some vehicles in a fleet may have experienced a different terrain, may not have executed certain operations at all.

- The second reason for this change in behavior of different vehicles is mainly due to the different external factors that the vehicle experiences when on road. Some example of these external factors are driver performance, different vehicle models in the same fleet, etc.
3.2.2.3 **Silhouette index based cluster grouping.**

Once the clusters are generated for each bus, next step in this process is to identify clusters in two different buses that are most similar. The most similar clusters are also the clusters in the data space that has an high overlap area. To determine this overlapping clusters, silhouette index is used as a metric of evaluation. As shown in Figure 19 once the clusters are generated silhouette Index (SI) is computed between each cluster to all other clusters of the other vehicle. The result of SI metric is a value that ranges between 0 to 1, where 0 indicates overlapping clusters while 1 indicates a crisp separation exits between the clusters Figure 21. Once the SI is computed for all clusters of the other vehicle, a one dimensional array indicates the overlap of one cluster to all other clusters of a different vehicle. Percentile based outlier detection is used to evaluate this one dimensional array to detect the pair of clusters that have the largest overlap form the rest Figure 22. Hence at the end of this process there are clusters of time series data segments exhibiting similar behavior.

```
clusters of bus 369 [1.0, 5.0, 6.0, 2.0, 3.0, 7.0, 4.0, 5.0, 6.0]
clusters of bus 370 [0.0, 1.0, 2.0, 3.0, 4.0]

5.1 metric of cluster 0.0 from bus 370 to all the clusters from bus 369
0.641 0.631 0.365 0.185 0.059 0.696 0.977 0.979 0.974
[False, False, False, False, False, False, False, False, False]
cluster 0.0 from bus 370 is most similar to the cluster 2.0 of bus 369

5.1 metric of cluster 1.0 from bus 370 to all the clusters from bus 369
0.515 0.568 0.577 0.311 0.639 0.470 0.988 0.986 0.984
[False, False, False, False, False, False, False, False, False]
cluster 1.0 from bus 370 is most similar to the cluster 2.0 of bus 369

5.1 metric of cluster 2.0 from bus 370 to all the clusters from bus 369
0.197 0.245 0.778 0.628 0.671 0.45 0.984 0.987 0.985
[True, False, False, False, False, False, False, False, False]
cluster 2.0 from bus 370 is most similar to the cluster 1.0 of bus 369

5.1 metric of cluster 3.0 from bus 370 to all the clusters from bus 369
0.612 0.473 0.727 0.576 0.897 0.439 0.971 0.976 0.972
[False, False, False, False, False, False, False, False, False]
cluster 3.0 from bus 370 is most similar to the cluster 7.0 of bus 369

5.1 metric of cluster 4.0 from bus 370 to all the clusters from bus 369
0.856 0.811 0.877 0.813 0.642 0.76 0.966 0.968 0.968
[False, False, False, False, False, False, False, False, False]
cluster 4.0 from bus 370 is most similar to the cluster 3.0 of bus 369

cluster groups with similar behavior between buses 369 & 370 are ['03', '11', '12', '20', '21', '35']
```

Figure 21: Output of peer clusters with similar behavior based on SI metric.

3.2.2.4 **Class ratio metric**

As indicated above, given 2 vehicles bus A (with m clusters) and bus B (with n clusters), after the Affinity propagation clustering is applied,
SI metric is computed for mxn times. This creates an high computational complexity. In this step we evaluate each cluster based on the supervised operation highway in-city classification. The main focus of this step is to use supervised classifier results to reduce the computational complexity. Instead of computing SI metric between all the clusters, we propose finding similar clusters for clusters performing similar operation. To find clusters performing similar operation, a
class ratio metric is computed. class ratio metric value ranges from 0 to 1, lower value indicates the cluster contains more highway segments, and higher value indicates more in-city segments. Based on this metric, probable clusters are proposed. As show in figure all the proposed clusters contain the cluster pair as seen before using this metric. As also seen in the test case, this metric provides correct probable clusters for SI metric to be computed in the further steps. Thus reducing the computational complexity from 45 computations to 28 computations Figure 23.

Figure 23: Peer cluster suggestion based on CI ratios.
DISCUSSION OF RESULTS.

The main goal of the thesis is to recognize the time instances when 2 different buses in a fleet have similar behavior. Given the time segments of buses in a fleet exhibiting similar behavior, fault detection based on anomaly deviation becomes more efficient. To achieve this goal, both supervised and unsupervised approaches were explored. In supervised approach different classifiers were trained to predict highway or in-city operations accurately, after many preprocessing steps as discussed in previous section 3.1.

To further investigate the existence of other operations and more efficient feature set, an unsupervised approach of clustering, different combinations of features and clustering algorithm results were implemented. From this unsupervised approach we generated a V-measure matrix that portrays the similarity of each clustering result to the rest. If there really existed only 2 unique operations(highway and in-city), we would have a data space of one cluster for each of the 3 different clustering algorithms, whereas the data space of clustering results of different algorithms contained multiple clusters for each algorithm as seen in the Figure 13 and Figure 12. Also exploring the most dense clusters for each clustering algorithm in Figure 13 backtracking to the cluster center data point would give the feature set of the clustering result to the rest in that gaussian fit. Evaluation of such feature space from most similar clustering result to the least similar clustering results for each clustering algorithm was conducted. The results as shown in Table 2 and Table 3 indicate better performance in both supervised classifiers and unsupervised clustering technique with the proposed feature set.

In unsupervised approach, apart from exploring for existence of other operations and identifying consistent feature set, we also propose a data driven method to group time segments from different buses exhibiting similar behavior. In this approach we focus on identifying different clusters that exist in the data space of 2 different buses and evaluating the similarity of behavior based on silhouette index. Affinity propagation clustering technique is used for identifying the number of clusters present in the data space.

Initially, silhouette index(SI) is computed for each cluster from one bus to all clusters in the other bus to identify the overlapping clusters. The intuition assumption of clusters having similar operation(highway and in-city) segments is more likely to be exhibiting the same behav-
ior is evaluated by proposing probable clusters with similar percent of operation segments (Class ratio metric). For each proposed cluster to be overlapped based on the class ratio metric was verified to contain the cluster with highest overlap. Thus lowering the computational complexity by computing silhouette index for each cluster from one bus to the clusters proposed based on class ratio metric. To validate this unsupervised approach another test case indicating the above described methodology is show in the next section.

4.1 Feature space exploration for vehicle dataset

To implement the feature evaluation technique as discussed earlier in Chapter 3, v-matrix for the vehicle dataset is computed and the PCA reduced data space of v-matrix is shown in Figure 24.

![Different clustering results of various feature space](image)

Figure 24: PCA reduced v-matrix for vehicle dataset.

As indicated in Figure 25 feature 1 (engine speed) is the most important feature since it has the highest silhouette index compared to the rest.

4.2 Test Case

In unsupervised approach clusters are generated after PCA dimensionality reduction as shown in Figure 26 and Figure 27 for both buses 369 and 372. Since the dimension space of both PCA clusters
(a) Silhouette index for feature 0 is 0.009
(b) Silhouette index for feature 1 is 0.313
(c) Silhouette index for feature 2 is -0.004
(d) Silhouette index for feature 3 is -0.003
(e) Silhouette index for feature 4 is 0.278
(f) Silhouette index for feature 5 is -0.004
(g) Silhouette index for feature 6 is 0.021

Figure 25: v-matrix based feature evaluation on vehicle dataset

[December 4, 2017 at 7:23 – classicthesis ]
are different, to generate peer groups of similar clusters, the original data space is used to find the clusters that overlap and the clusters that are far apart.

![Affinity propagation clustering of the bus 369](image)

Figure 26: Affinity propagation clustering of bus 369.

Silhouette index (SI) is computed for each cluster from bus 369 to all the clusters of bus 372. As observable in Figure 28, there are 9 clusters in bus 369 and 6 clusters in bus 372. Initially SI metric is computed for the cluster with label ‘1.0’ from bus 372 to all other sisters in bus 369. Thus generating a 1 Dimensional array of values that usually tend to lie near the value 0.9 except for 1 entry. The cluster ‘1.0’ has a SI metric of 0.4 against cluster ‘0.0’ of the other bus. This is a very low score than the rest of the other comparisons, indicating that the cluster ‘1.0’ of bus 372 is most similar to cluster ‘0.0’ of the bus 369. to automate this process of finding clusters that overlap iteratively, percentile based outliers is used to detect the SI metric of the clusters that are different from the rest for the clusters below a certain threshold. Figure 29.

Thus iterating the above across different clusters, groups of clusters that have similar behavior are realized as shown in Figure 28.

To reduce the computational complexity of comparing each cluster of one bus to all other clusters of the other bus, a class ratio metric is generated for each cluster using the supervised classifier prediction.
Based on this class ratio metric, clusters with similar operations are proposed in Figure 30. As it can be observed, initially acquired clusters group from Figure 28 is present in each of the proposed cluster with similar operation in Figure 30. Thus reducing the initial computations to 30.
**Clusters of bus 369**: [1.0, 0.0, 0.0, 2.0, 3.0, 7.0, 4.0, 5.0, 6.0]

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distance</th>
<th>Silhouette</th>
<th>Similarity</th>
<th>Predication</th>
<th>Cluster 1.0 is most similar to the cluster 0.0 of bus 369</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.928</td>
<td>0.914</td>
<td>0.338</td>
<td>False, False, True, False, False, False, False, False, False, False, False, False, False, False, False, False</td>
<td></td>
</tr>
</tbody>
</table>

**Clusters of bus 372**: [1.0, 2.0, 4.0, 0.0, 5.0, 3.0]

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distance</th>
<th>Silhouette</th>
<th>Similarity</th>
<th>Predication</th>
<th>Cluster 2.0 is most similar to the cluster 0.0 of bus 369</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0.623</td>
<td>0.625</td>
<td>0.337</td>
<td>False, False, True, False, False, False, False, False, False</td>
<td></td>
</tr>
</tbody>
</table>

**Clusters of bus 372**: [1.0, 2.0, 4.0, 0.0, 5.0, 3.0]

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distance</th>
<th>Silhouette</th>
<th>Similarity</th>
<th>Predication</th>
<th>Cluster 4.0 is most similar to the cluster 1.0 of bus 369</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>0.208</td>
<td>0.309</td>
<td>0.677</td>
<td>True, False, False, False, False, False, False, False, False, False, False, False, False, False, False, False</td>
<td></td>
</tr>
</tbody>
</table>

**Clusters of bus 372**: [1.0, 2.0, 4.0, 0.0, 5.0, 3.0]

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distance</th>
<th>Silhouette</th>
<th>Similarity</th>
<th>Predication</th>
<th>Cluster 0.0 is most similar to the cluster 3.0 of bus 369</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.813</td>
<td>0.74</td>
<td>0.772</td>
<td>False, False, False, True, False, False, False, False, False, False, False, False, False, False, False, False</td>
<td></td>
</tr>
</tbody>
</table>

**Clusters of bus 372**: [1.0, 2.0, 4.0, 0.0, 5.0, 3.0]

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distance</th>
<th>Silhouette</th>
<th>Similarity</th>
<th>Predication</th>
<th>Cluster 5.0 is most similar to the cluster 8.0 of bus 369</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>0.219</td>
<td>0.018</td>
<td>0.793</td>
<td>False, True, False, False, False, False, False, False, False, False, False, False, False, False, False, False</td>
<td></td>
</tr>
</tbody>
</table>

**Clusters of bus 372**: [1.0, 2.0, 4.0, 0.0, 5.0, 3.0]

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distance</th>
<th>Silhouette</th>
<th>Similarity</th>
<th>Predication</th>
<th>Cluster 3.0 is most similar to the cluster 7.0 of bus 369</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>0.535</td>
<td>0.305</td>
<td>0.7</td>
<td>False, False, False, False, True, False, False, False, False, False, False, False, False, False, False, False</td>
<td></td>
</tr>
</tbody>
</table>

Cluster groups with similar behavior between buses 369 & 372 are [12, 20, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70].

Figure 28: Affinity propagation and Silhouette index based peer group recognition.
Figure 29: Percentile based outlier detection of Silhouette index metrics.
cluster 1.0 of bus 372 is similar to the following clusters of 369:
cluster 1.0
cluster 0.0
cluster 2.0
cluster 1.0 of bus 372 is similar to the following clusters of 369:
cluster 1.0
cluster 0.0
cluster 2.0
cluster 4.0 of bus 372 is similar to the following clusters of 369:
cluster 1.0
cluster 0.0
cluster 2.0
cluster 0.0 of bus 372 is similar to the following clusters of 369:
cluster 8.0
cluster 2.0
cluster 3.0
cluster 7.0
cluster 4.0
cluster 5.0
cluster 6.0
cluster 5.0 of bus 372 is similar to the following clusters of 369:
cluster 8.0
cluster 2.0
cluster 3.0
cluster 7.0
cluster 4.0
cluster 5.0
cluster 6.0
cluster 3.0 of bus 372 is similar to the following clusters of 369:
cluster 8.0
cluster 2.0
cluster 3.0
cluster 7.0
cluster 4.0
cluster 5.0
cluster 6.0

Figure 30: Proposed clusters based on class ratio metric.
Multivariate time series data acquired from different buses in fleet, performing many operations have been explored in both supervised and unsupervised techniques. Adaptive segmentation is implemented to reduce data representation and these segments are used as input to supervised and unsupervised approach. With the help of GPS signals labeled data were generated. These labeled data based on the GPS timestamps were used to train and validate different classifiers. In supervised approach, standard classifiers were trained to predict adaptive time segments as highway and in-city operations. Since such ground truth or labeled data aren’t available for other operations, unsupervised approach was implemented for feature space exploration and also for peer group recognition based on Affinity propagation clustering and silhouette index.

The main focus of this thesis was achieved to generate time segments that exhibit similar vehicle performance, by either identifying the operation of the vehicle in supervised approach or by peer group recognition of clusters based on vehicle behavior in unsupervised approach. v-matrix based feature space exploration was verified on synthetic dataset. Also, feature space exploration led to the discovery of consistent feature and multiple clusters that exhibit the existence of other significant operations. Both supervised and unsupervised approaches were explored, evaluated and also combined to recognize peer groups of clusters in lower computations. An interesting approach towards visualizing the feature space to provide optimal combination of feature set that would improve the performance of both supervised and unsupervised techniques were evaluated.

5.1 Future Work

There are a wide range of tasks that can yield interesting results as discussed below:

- Future work would include exploring feature space evaluation with different combination of threshold error for adaptive segmenting, different clustering algorithms and other parameters to be optimized.

- Collecting more data such as traffic density around the bus, to have ground truth for exploring and validating other orthogonal operations of vehicle that exist.
• Exploring different techniques to understand and evaluate the influence of existing orthogonal and hierarchical operations on performance of a vehicle.

• Evaluation of impact in implementing both supervised and unsupervised techniques discussed in this thesis for fault detection.

• One more important and effective future work would be in aggregating the labels of time series segments based on the confidence in current predicted label and the labels of its adjacent time segments. Also to explore the time segments in finding the optimal time instance of transition within different operations.


[December 4, 2017 at 7:23 – classicthesis]


DECLARATION

I hereby declare that the work presented in this document is true to the best of my knowledge.

Halmstad, April 2017

Karthik Bangalore
Girijeswara, December 4,
2017