

Knowledge Extraction from Real-World Logged Truck Data

Thomas Grubinger¹ and Nicholas Wickström
Halmstad University

Anders Björklund and Magnus Hellring
Volvo AB

Copyright © 2009 SAE International

ABSTRACT

In recent years more data is logged from the electronic control units on-board in commercial vehicles. Typically, the data is transferred from the vehicle at the workshop to a centralized storage for future analysis. This vast amount of data is used for debugging, as a knowledgebase for the design engineer and as a tool for service planning.

Manual analysis of this data is often time consuming, due to the rich amount of information contained. However, there is an opportunity to automatically assist in the process based on knowledge discovery techniques, even directly when the trucks data is first offloaded at the workshop. One typical example of how this technique could be helpful is when two groups of trucks behave differently, e.g. one well-functioning group and one faulty group, when the two groups have the same specification. The desired information is the specific difference in the logged data, e.g. what particular sensors or signals are different.

An evaluation cycle is proposed and applied to extract knowledge from three different large real-world data-sets measured on Volvo long haulage trucks. Information in the logged data that describes the vehicle's operating environment, allows the detection of trucks that are operated differently from their intended use. Experiments to find such vehicles were conducted and recommendations for an automated application are given.

INTRODUCTION

The increasing amount of data that is logged by electronic systems of commercial vehicles opens new opportunities to vehicle manufacturers. The vehicle data itself is commonly logged by the various electronic control units, which log, e.g., information from the engine control and air suspension. Despite the common use of data for debugging, a proper analysis of the data can provide a knowledge base, which can assist in vehicle design and service planning.

SCOPE AND MOTIVATION The evaluation process proposed in this work neither intends to replace nor to minimize current efforts currently done in design- or quality control experiments being conducted in laboratory structures. The analysis of the vehicle data clearly will not cope with the accuracy of experiments that were well planned and conducted using special equipment. For example Karlsson [1] discusses the influence of parameters like lateral loads, transport mission and the driver's impact on the fatigue life of a truck. Another paper by Karlson [2] evaluates the characterizations of road curves in a driver-independent way. Two papers by Öijer and Edlund [3, 4] deal with the influence of road conditions and transient obstacles of the road on the durability of a vehicle.

The aim is to gain additional insights, by not only analyzing a relatively small group of vehicles, which can be done more efficiently with special measurement equipment, but to analyze lots of vehicles that are operated all around the world, utilized in multiple forms and different operating environments.

¹During the time this work was conducted Thomas Grubinger was with Halmstad University. He is now with the Medical University of Innsbruck, Austria

The focus of the evaluation framework proposed is to compare different populations of vehicles. The dissimilarity can be of various forms; e.g., two data-sets used in this work were chosen from populations that differed in their operating environment by road conditions or vehicle load. Other options include populations that are utilized with different configurations or groups of vehicles that are known to behave differently.

APPLICABILITY AND DESIGN OF EXPERIMENTS Basically the proposed methodology can be used if two groups of vehicles differ by any relevant factor, including different operating environments, driving styles, usages and configurations. However, the evaluation probably only yields good results when each group is at least relatively homogeneous. This is due to that the variables logged by the vehicles' systems are influenced by various properties, including dissimilarities in the operating environments, varying driving styles as well as differences in the vehicles' configurations.

This implies that an exact measurement of e.g., the influence of one factor, e.g., a different topography is not possible. However data-sets can be arranged to minimize the influence of factors, except the factor under study. Thus populations should be chosen to have different properties, e.g., a different topography, but as similar as possible to other influencing factors. Besides similar operating environments and similar vehicle configurations, the origin of the data was found to be important. This is not only because of the dissimilar operating environments in the different regions, but also because of dissimilarities in use, e.g., different speed limits and varying driving habits in different cultures.

POSSIBLE AREAS OF APPLICATIONS One big area of applications are groups of vehicles that are known to be operated differently, e.g., in different countries, operating environments or in different cultures. In these applications the dissimilarities in the data can give insight on the different loads that are put on the vehicles' subsystems. This knowledge in turn gives valuable feedback for the design of vehicles and the planning of service intervals. Service intervals can be tailored more precisely to customers' requirements to maximize vehicle reliability and minimize service costs. As an example, different use of the vehicles in markets and regions can be used to customize vehicles accordingly to maximize efficiency.

When groups of vehicles with the same specification behave differently the proposed evaluation cycle can assist in finding data that systematically differ among these groups, e.g., one faulty and one well-functioning group. Since the logged data corresponds to a particular sensor of the vehicle, insight on the causes of the different behavior is gained.

In cases where vehicle populations are configured differently the analysis of the data can give insight on the impact of this difference, e.g., if one configuration results in a higher load on some subsystems. One example is a vehicle sold in two engine configurations. The more

powerful configuration might put more load on some parts of the vehicle, which can be verified by the data logged. A data-set corresponding to this particular example was evaluated during the work and is further discussed at the end of the *Examples of Evaluated Data Sets* section.

METHODOLOGY The algorithms applied were chosen to extract information from the data automatically, without incorporating any knowledge about its source. The algorithms used are well established and have performed well in different applications including medicine, biology, finance and marketing [5].

The methodology used in this work originates from the field of *machine learning*, a branch of *artificial intelligence* [6] that deals with "*learning*" from data, that is to extract patterns and trends from the data. The field of machine learning can roughly be divided into the two sub categories *supervised* and *unsupervised* learning. An introduction to machine learning algorithms is given in [5].

The field of **supervised learning** relies on the principle of learning by example. That means the outcome of the learning process is guided by the known output categories. When the task of a supervised learning duty is to assign a given input to a discrete output (*classes*), the task is also called a *classification task*. The goal of a classification task is to estimate the output categories based on the inputs, as precisely as possible.

On the other hand, **unsupervised learning** only uses the input variables without incorporating the output categories. Hence, the result will not necessarily be the best way to describe the different output categories, but will be the statistically most significant phenomenon in the data. As a result, unsupervised learning methods are less prone to over-fitting of the data. This makes it possible to extract dependencies and to gain knowledge from the data inherent in the given problem. Unsupervised methods where the results are discrete groups (*clusters*) are termed *clustering methods*.

This work focuses on unsupervised methods, as the focus was on extracting knowledge from the data in a general sense, rather than building an optimal tool for classification. To support the possibility of extracting information inherent in the problem (i.e., the factor under study), methods were incorporated, which do not need a pre-specified number of clusters in the output. Special attention was given to the visualization of the data, which made advanced conclusions possible about the vehicle populations.

EVALUATION FRAMEWORK

Figure 1 is an overview of the individual tasks, which were necessary for this work. It can be observed that the stages are similar to the *KDD*² [7] model, which is a more general illustration of the task.

²Knowledge Discovery from Databases

In the reminder of this subsection are short descriptions of the individual tasks. Figures displayed in this section show

values or have other undesirable properties that negatively influence the outcome of the clustering task.

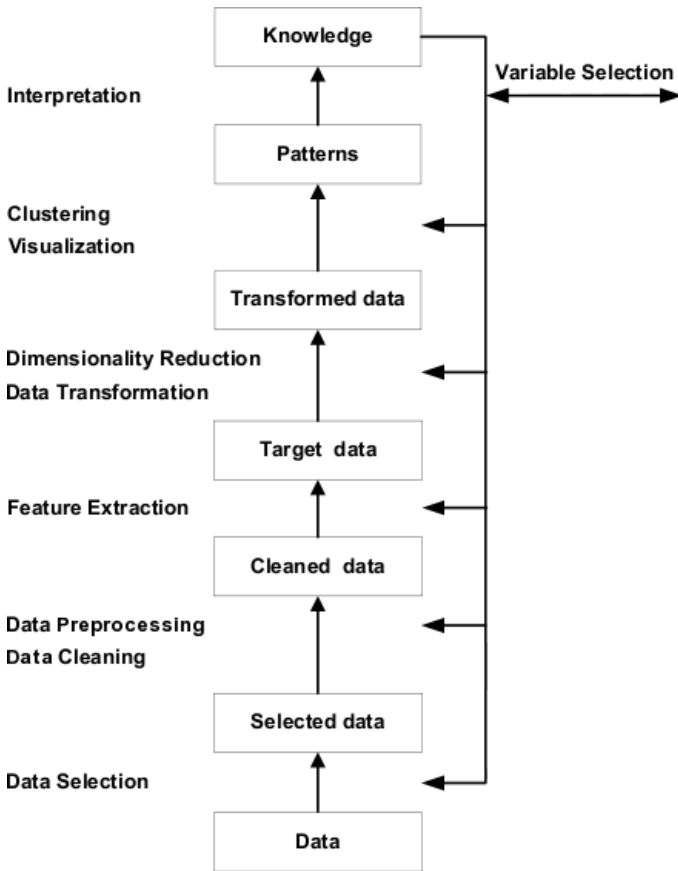


Figure 1: Overview of the tasks and the process for knowledge extraction proposed in this work

the results from different data-sets and are discussed in a general way. The interpretation of the figures will be covered in the subsequent section *Examples of Evaluated Data-Sets*.

DATA SELECTION An important step in every clustering task is to use an optimal subset of variables. Furthermore, a reduced number of inputs also facilitates a better understanding of the underlying process, improves the performance of the algorithm used and enables the visualization of data. When the number of variables is high it is not possible to evaluate all possible input combinations. The selection of the relevant variables was based on the clustering performance, using forward selection and backward elimination [8], as well as on the interpretation of the output from visualization techniques that were used.

The final model for the clustering task is usually not only chosen by selecting the inputs with the best performance, but also considering the models simplicity. This is because a model' with higher complexity may perform better on the data from which it was constructed. This issue is also known as the bias-variance tradeoff.

DATA CLEANING The task of *data cleaning* deals with the removal of observations which have either erroneous

Tasks necessary in the data cleaning step are the removal of vehicles with incomplete data as well as observation with low operating time. Observations with variables that have very different values compared to other vehicles are removed as well. These observations are called *outliers* and the task *outlier detection*. Outliers are removed because it is usually not wise to include observations that only rarely occur. On the one hand, such observations can be wrong measurements that would only add wrong information. Also, if the measurement is perfectly correct there are still only very few samples of what was measured, which usually leads to a poorer model [9].

The simplest approach is to remove observations that contain variables that are more than 3 or 4 standard deviations from their class mean. An overview of mathematically more sophisticated methods is given in e.g., [10, 11, 12].

PREPROCESSING THE VEHICLE DATA In this subsection the general issues, e.g., the format of the data, specific properties associated with it, as well as the preprocessing steps necessary, are discussed.

Data Format Vehicle data comprises variables that are arranged in various formats. Variants of these different formats are described below.

Single Variables Often log the time a vehicle was in a certain condition, e.g., how long the engine ran at idle.

Commonly, vehicles also log variables counting how many times a certain condition was reached, e.g., how often the clutch or the brake was pressed.

1D- Feature Vectors are single variables arranged in a vector, where each variable represents the time the vehicle has been in a certain condition.

An instance of this data format logs information concerning the speed of a vehicle. The first variable logs the time the vehicle has spent between 0 and 40 km/h, the second between 40 and 56 km/h and so on.

A 1D- feature vector can effectively be visualized as a histogram, as shown in Figure 2.

2D- Feature Vectors are organized similar to the 2D-feature vectors described above and are arranged in matrices. One example of data in 2D is how long a vehicle has spent under a combination of factors, e.g., engine load and engine speed.

An example of this data format logs information regarding torque and the speed of the trucks engine. One variable of this data holds the time the vehicle

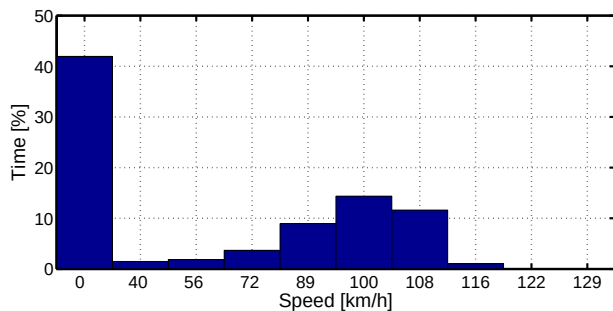


Figure 2: Example of a 1D- feature vector, interpreted as a histogram, that shows the time the truck has spent at a certain vehicle speed.

has spent between a certain range of torque and engine speed. An instance drawn from the example discussed, is interpreted and visualized in Figure 3 as a 3D- histogram.

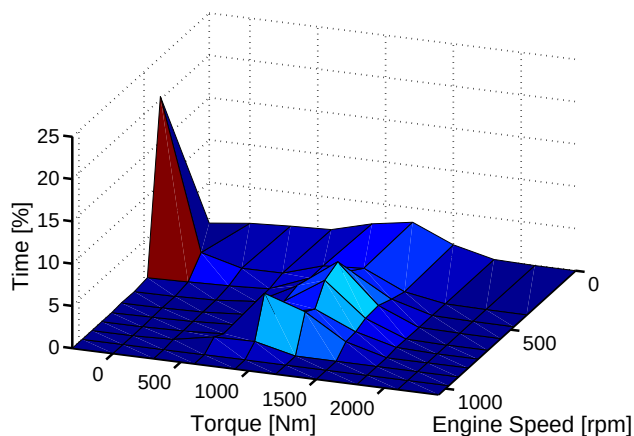


Figure 3: Example of a 2D- feature vector that show the time the vehicle has spent at a certain torque and engine speed.

Data Specific Issues and Pre-Processing From the format, as well as from the other specific characteristics of the data the following issues arise:

Normalization of Feature Vectors The values of the variables in the feature vectors depend on time. However, we are interested more in comparing the distributions of the instances of the feature vectors than the absolute values. The absolute values of a histogram coming from Truck A, have x times more operating time than Truck B. As a result the values of the data from Truck A, are on average x times higher than the values of the vector coming from Truck B. This issue can be easily solved by dividing each variable by the sum of all variables.

Idle Time is another specific property of the vehicle data, which was found to be of greatest importance

in truck data. By looking at Figure 2 the first variable comprises time when the vehicle is not moving. The histogram viewed originates from an American long haulage truck. Commonly American truck drivers sleep in their trucks, while the engine is running. This is the reason why the histogram shows the vehicle is operated at idle or very low speed for such a long period of time.

An instance of this problem is shown in Figure 3, where the maximum value is at around an engine speed of 800 rpm and 0 Nm torque, thus the condition when the trucks engine runs at idle. For such data the variables describing the idle time should be removed before the variables are divided by their sum.

FEATURE EXTRACTION This step deals with the extraction of new variables from already existing variables, which are then incorporated in the set of possible inputs. An issue that had to be addressed before working on the individual tasks was how the information held by the instances of the feature vectors could be represented best. One way is to use the values of the individual variables directly. However, there are other possibilities to extract additional variables from the data arranged in a vector format, which describe the shape of the histogram. In the case of the 2D-feature vector, two 1D-feature vectors were calculated by summing up the values on the horizontal and on the vertical axis respectively.

There are a number of statistical features (e.g., Variance, Skewness, Kurtosis, Median, lower and upper quartiles) that can be easily calculated from feature vectors.

An additional possibility is to do regression analysis on the data vectors and use the polynomial coefficients as input for the learning task. During this work polynomials of order one and two have been successfully applied.

DATA TRANSFORMATIONS For most classification and clustering methods it makes a big difference in performance when the data is transformed in an appropriate way before it serves as an input. It is important that all variables are of the same size such that they are considered equally important by the different algorithms. A simple way to achieve this is called *standardization* [9], which is a method that transforms the input variables to zero mean and unit variance.

DIMENSIONALITY REDUCTION The idea of using techniques for *dimensionality reduction* is to use only a small subset of variables of the given input space such that the important information is preserved. This is an essential step, since most algorithms used in machine learning work better on low dimensional data, which is mainly due to properties of high dimensional data referred to as the *curse of dimensionality* [5, 9].