

A Load Spectrum Data based Data Mining System for Identifying Different Types of Vehicle Usage of a Hybrid Electric Vehicle Fleet

Philipp Bergmeir
Esslingen University of Applied Sciences

Christof Nitsche
Daimler AG

Jürgen Nonnast
Esslingen University of Applied Sciences

Michael Bargende
University of Stuttgart

ABSTRACT

In order to achieve high customer satisfaction and to avoid high warranty costs caused by component failures of the power-train of hybrid electric vehicles (HEV), car manufacturers have to optimize the dimensioning of these elements. Hence, it is obligatory for them to gain knowledge about the different types of vehicle usage being predominant all over the world. Therefore, in this paper we present a Data Mining system that employs a Random Forest (RF) based dissimilarity measure in the dimensionality reduction technique t-Distributed Stochastic Neighbor Embedding (t-SNE) to automatically identify and visualize different types of vehicle usage by applying these methods to aggregated logged on-board data, i.e., load spectrum data. This kind of data is calculated and recorded directly on the control units of the vehicles and consists of aggregated numerical data, like the histogram of the velocity signal or the traveled distance of a vehicle. We empirically demonstrate the performance of the proposed Data Mining system by carrying out a real-world case study using load spectrum data of a HEV fleet, containing the logged on-board data of 6670 passenger cars. The results show that the system is able to learn from approximately 700 variables, which describe the usage of each vehicle, a two- or three-dimensional embedding, making it possible to visualize the different types of vehicle usage.

CITATION: Bergmeir, P., Nitsche, C., Nonnast, J., and Bargende, M., "A Load Spectrum Data based Data Mining System for Identifying Different Types of Vehicle Usage of a Hybrid Electric Vehicle Fleet," *SAE Int. J. Alt. Power*. 5(1):2016, doi:10.4271/2016-01-0278.

INTRODUCTION

Nowadays, modern vehicles are equipped with several control units, such as an engine control unit (ECU) or a battery management system (BMS), where different types of data are recorded, e.g., data describing the usage or loads of the vehicle. Engineers use these kinds of data, amongst others, to optimize the dimensioning of the power-train and its components or to identify failures in the operating mode of the vehicle.

Moreover, the fast development of powerful data collection and storage tools makes it now possible to record a huge amount of onboard data of each vehicle. However, uncovering valuable information, e.g., different patterns of vehicle usage, and transforming it into structured knowledge becomes more and more difficult [1].

In this paper, we investigate the usability of several dimensionality reduction techniques, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) [2] and Isometric Feature Mapping (Isomap) [3],

to create a two- or three-dimensional embedding of the high-dimensional logged on-board data that are recorded for a HEV fleet, consisting of 6670 passenger cars. Furthermore, we propose a new variant of t-SNE that employs a Random Forest based dissimilarity measure [4] instead of a Euclidean distance and show its superiority on the studied dataset.

Thereby, the considered on-board data exclusively consists of load spectrum data, i.e., counting data that are usually employed for calculating the fatigue life of components in fatigue analysis. Since the aggregation of stress-time functions to load spectrum data leads to a loss of information [5], we also aim to find out whether this data transformation preserves enough information to visualize different types of vehicle usage. Furthermore, we want to learn a low-dimensional representation of each vehicle, i.e., of its load spectrum data, not by only incorporating a pre-selected subset of these load spectra, like only the histogram of the vehicle speed or the level

distance, but by using all of its available load spectrum data. Thus, we hope to achieve a better two- or three-dimensional representation of the overall vehicle usage.

The structure of the remainder of this paper is organized as follows. We complete the first section with a short survey of related work, before we briefly introduce load spectrum data which build the basis of our analysis in the second section. In the third section, we explain the dimensionality reduction technique t-SNE in detail and describe how the Random Forest based dissimilarity measure is computed and incorporated in this method. Then, the fourth section presents how the data is preprocessed and what parameter settings are chosen. The fifth section provides the results and compares them to the outcome of several other frequently used dimensionality reduction methods. Finally, the sixth section summarizes the paper and provides a brief outlook on future work.

Related Work

There are only a few publications about applying Data Mining or Machine Learning algorithms to load spectrum data, although in the automotive industry there are a lot of logged on-board data available, nowadays.

In our recent publications [6], [7] and [8], we empirically demonstrate the usability of several Machine Learning techniques, such as Random Forests [9], when applied to load spectrum data that are recorded for a huge HEV fleet, in order to distinguish between non-faulty HEV and vehicles suffering from a failure of the hybrid car battery. Moreover, we show that the proposed framework is able to select a small set of load spectrum classes that can be related to battery failures.

In [10], a case-study is presented where prognostic information on remaining useful life (RUL) of lead-acid batteries in Scania heavy-duty trucks is computed by applying Random Survival Forests to maintenance and logged on-board vehicle data, i.e., load spectrum data.

Prytz et al. [11] describe how random forest models can be used for estimating the probability that air compressors in commercial trucks and buses survive until the next planned service stops on the basis of logged on-board data.

In summary, in spite of the availability of large amounts of data describing the usage and loads of a vehicle, there is still a lack of publications that investigate existing or develop new methods to extract useful knowledge from this kind of data. In order to be able to improve, amongst others, the reliability and the carbon footprints of future vehicles, it will become inevitable to spend more and more time on research regarding these topics.

LOAD SPECTRUM DATA

During their lifetime, mechanical and electrical components, such as the parts belonging to the power-train of a HEV, are subject to different kinds of stress of distinct strength. In order to learn which

loads occur in operating mode as well as in resting state and what effects they have on the considered component, engineers usually measure and record loads under real-world operating conditions [5].

Since the fatigue life of a component depends, amongst others, on the magnitude of the amplitude of alternating stress and its frequency, over the last decades many counting methods have been developed that transform stress-time functions to frequency distributions. Thereby, some characteristics of the stress-time functions, e.g., the chronology of the occurrence of different events, get lost. Hence, it has to be decided separately for each use case whether this data reduction step is feasible or not [5].

These counting methods are often categorized in “one-” and “two-parameter” counting methods, respectively, depending on the number of signals they process concurrently. However, most of these methods could also be generalized to transform more than two signals. All of them result in a so-called load spectrum. Sometimes the result of two-parameter counting is called a load matrix, but since we do not distinguish between load spectrum and matrix in our analysis, we only use the denomination “load spectrum” [5].

The majority of our data results from a counting method called “level distribution counting” [5]. Therefore, we explain it briefly in the following. Applying “level distribution counting” to a signal means splitting the value range in disjunctive classes of interest, e.g., intervals. The defined classes do not necessarily have to be of the same width. Afterward, the whole measurement time is cut into equidistant time intervals and at each of the resulting points in time the signal’s current value is queried [7]. Then, the counter of the interval to which the value belongs to is increased by one.

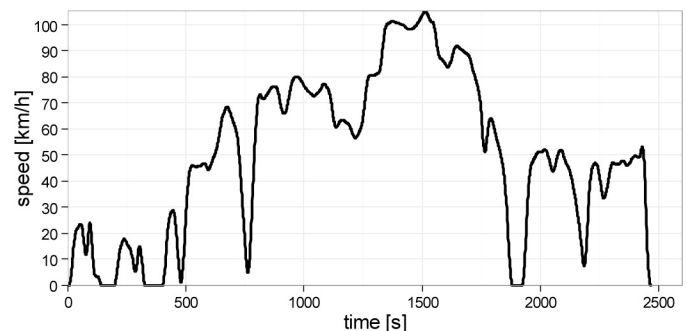


Figure 1. Example of a stress-time function of the signal “speed” [6]

For illustration purposes, Figure 1 shows an instance of a stress-time function of the signal “speed” and Figure 2 visualizes the corresponding load spectrum that results from a one-parameter level distribution counting applied to the signal. Thereby, the speed range is divided into the twelve intervals 0-10, 10-20, ..., 110-120km/h and the value of the signal is queried at every second.

In this work, we analyze the load spectrum data of 6670 HEV which define a subset of the data used in [6], [7] and [8]. The 35 considered load spectra are computed directly on the control units, e.g., on the ECU, of each vehicle, i.e., no raw signals are available. They result from transforming heterogeneous signals that are recorded during real driving, like the temperature of the hybrid car battery or the current

of the electric machine. Additionally, they contain simple event counters, such as the number of engine starts, and vary in the number of intervals the signals are divided into. Moreover, we perform different calculations on the existing load spectra, e.g., summing up all class entries of a load spectrum, to derive 17 additional, “synthetic” load spectra. Furthermore, we normalize each load spectrum that contains more than one class to sum one, i.e., we determine for each class the relative frequency. However, simple counters, like the traveled distance, are not normalized. Finally, we only consider vehicles of countries in which for at least 25 cars the load spectrum data is available in our database.

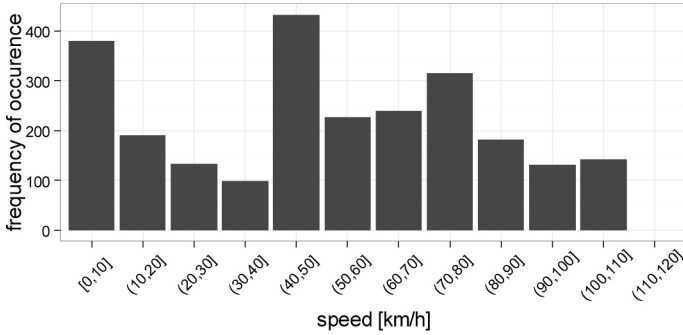


Figure 2. Visualization of the load spectrum “speed” resulting from a one-parameter level distribution counting applied to the signal shown in Figure 1 [6]

DIMENSIONALITY REDUCTION METHODS

Dimensionality reduction techniques are often used to project high-dimensional data into a low-dimensional, e.g., two- or three-dimensional space, to facilitate the visualization of the data, e.g., to be able to create a scatterplot. Thereby, the main goal of these methods is to find a low-dimensional data representation that preserves as much significant structure of the high-dimensional data as possible [2].

Generally, the existing techniques can be divided into two main groups: linear and nonlinear reduction methods. The aim of the traditional, linear approaches, like Principal Components Analysis (PCA) [12] and classical Multidimensional Scaling (MDS) [13], [14], is to model dissimilar data objects far apart in the low-dimensional space, while nonlinear techniques, such as Sammon Mapping [15], Isomap [3], and Locally Linear Embedding (LLE) [16], focus on keeping very similar data points close together. However, the majority of both groups of methods are not able to preserve both the local and the global structure of the data in the low-dimensional data representation. In order to overcome this problem, t-SNE [2] has been developed by van der Maaten and Hinton in 2008. Since this method is the core element of our proposed approach, we describe it in detail in the following section.

t-Distributed Stochastic Neighbor Embedding

In recent studies, t-SNE has been successfully used to determine two-dimensional projections of high-dimensional data, e.g., mouse brain data [17] and metagenomics data [18]. It is an extension of the formerly developed Stochastic Neighbor Embedding (SNE) [19] that

alleviates the optimization of the objective function of SNE. Moreover, it tries to circumvent a problem the authors refer to as the “crowding problem” [2], which we do not discuss here.

Given a dataset of high-dimensional input objects $X = \{x_1, x_2, \dots, x_N\}$, where $x_i \in \mathbb{R}^p$, t-SNE determines a low-dimensional data representation $Y = \{y_1, y_2, \dots, y_N\}$ called a “map” that is formed by the so-called “map points” $y_i \in \mathbb{R}^m$, where $m \ll p$. Typical choices for m are 2 or 3, respectively. The main goal of t-SNE is to preserve as much of the local structure of the high-dimensional data as possible, while also capturing the global structure like the presence of clusters. Thereby, it measures the pairwise similarity between two arbitrary data objects x_i and x_j in the high-dimensional space X by the following joint probability p_{ij} that symmetrizes two conditional probabilities [20]:

$$p_{j|i} = \frac{\exp(-d(x_i, x_j)^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-d(x_i, x_k)^2 / 2\sigma_i^2)}, \quad p_{i|i} = 0, \quad (1)$$

$$p_{i|j} = \frac{\exp(-d(x_j, x_i)^2 / 2\sigma_j^2)}{\sum_{k \neq j} \exp(-d(x_j, x_k)^2 / 2\sigma_j^2)}, \quad p_{j|j} = 0, \quad (2)$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \quad (3)$$

where σ_i and σ_j designate the bandwidth of the Gaussian kernels and $d(x_i, x_j)$ denominates the pairwise distance or dissimilarity between x_i and x_j , e.g., the Euclidean distance $d(x_i, x_j) = \|x_i - x_j\|_2$. In the case of $d(x_i, x_j)$ being the Euclidean distance, the conditional probability $p_{j|i}$ (cf. Eq. (1)) captures the probability that x_i selects x_j as its neighbor when neighbors are picked in proportion to their probability density under a Gaussian distribution centered at x_i .

The value of σ_i is determined separately for each data object x_i , since it is not likely that there is a single value of σ_i that is optimal for all data points in X . In fact, the density of the data is likely to vary, i.e., it might contain both dense and sparse regions, where a smaller value of σ_i is usually more appropriate in dense than in sparse regions. Therefore, σ_i is determined by a binary search such that the perplexity, $Perp(P_i)$, of the conditional distribution P_i that is induced by σ_i over all of the other data objects equals a fixed value specified by the user. The perplexity can be interpreted as a smooth measure of the effective number of neighbors and the performance of t-SNE is usually fairly robust to changes of this parameter [2]. Its recommended values lie in the range of 5 to 50 and it is formally defined, as follows:

$$Perp(P_i) = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}}. \quad (4)$$

In the m -dimensional embedding Y the similarity q_{ij} between two points y_i and y_j , i.e., between the corresponding map points of the data objects x_i and x_j , is measured using a heavy-tailed Student's t-distribution with a single degree of freedom: