Combination of Data-driven Feature Selection Methods with Domain Knowledge for Diagnosis of Railway Vehicles

Bernhard Girstmair[1], Andreas Haigermoser[2], and Justinian Rosca[3]

[1,2]*SIEMENS AG Österreich, Graz, Styria, 8020, Austria*

bernhard.girstmair@siemens.com
andreas.haigermoser@siemens.com

[3]*SIEMENS Corporate Technology, Princeton, NJ, 08540, USA*

justinian.rosca@siemens.com

ABSTRACT

Railway vehicles are generally maintained preventively within certain time periods. Condition based predictive maintenance strategies have a great economic potential so that modern trains are equipped with many sensors in order to perform diagnostics and prognostics of components. Methods for fault detection need appropriate feature subsets in order to achieve small in-sample and out-sample errors. In our case the typical feature selection approach using pure data-driven methods is difficult, as the number of possible feature sets is very large. On the other hand there exists rich domain knowledge and detailed physical models of the mechanical system. The aim is to combine this knowledge with the often used mathematical methods for feature selection for improving classification of cases when a faulty damper is present. Based on the dynamic equations of motion, this paper presents heuristic feature selection via the analysis of transfer functions. We describe several well-known methods of automated feature selection and a workflow which combines domain knowledge with automated methods. Results show that it is difficult to define features based only on domain-knowledge, but in combination with data-driven techniques good classification performance can be achieved.

1. INTRODUCTION

The investigated mechanical system is a conventional railway vehicle, which typically consists of four wheelsets, two bogies (leading and trailing bogie) and a car body. Within the bogie there are many mechanical parts such as dampers and springs that are generally maintained preventively within certain time periods. Railway vehicles run for a long time (up to thirty years) and about one-third of lifecycle costs is due to maintenance (Baumgartner, 2001). Condition based predictive maintenance strategies offer the possibility of large savings as well as improvement in reliability. Therefore modern trains are equipped with a large number and variety of sensors for diagnostics and prognostics of components of railway vehicles.

In our application a railway vehicle is equipped with up to 50 sensors. Most of them are acceleration sensors. In addition, process data from other subsystems is available. Sensors are mounted at several different positions and at different levels of suspension. To ensure good ride comfort there are usually two levels of suspension within a railway vehicle. Each level of suspension consists of several components such as coil springs, dampers, air springs and rubber elements. If one component of the system is faulty the dynamics of the railway vehicle changes. The feature selection process should find the most informative (leading to better human interpretations) and accurate (leading to smallest errors) sets of signatures or fingerprints computable from sensor and control data of the mechanical system.

Feature extraction converts the initial raw signals into more informative signatures of the system, while reducing the dimensionality of the input data. Many types of features can be defined, such as time domain, frequency domain, and time-frequency features. Kimotho and Sextro (2014) give a good overview of possible features. Typical features with a physical interpretation in our application are standard deviation and maximum value over time frames within defined frequency bands. The reduced representation of the data obtained with feature extraction, and further with the selection of a subset of relevant features is expected to contain sufficient information for diagnosing or predicting the health state of the system or its components with statistical methods such as classification, abnormality detection, clustering and regression. Therefore it is important that selected features contain as much information as possible regarding the system states of interest, while removed features and data do not incur an information loss. Selecting the most appropriate features is of paramount importance within the process of designing a high performance classifier. Feature selection is a well-known problem in machine learning and has generated a large volume of research. Unfortunately feature selection methods are computationally very costly. This paper takes the perspective of combining domain knowledge and statistical methods to address the feature selection problem and demonstrates that both accurate and informative results can be achieved in diagnostics and prognostic problems for train vehicles.

The layout of the paper is as follows. Section 2 presents the physics-based principles behind this work, covering the dynamic equations of motion, physics-based simulation and principles for heuristic feature selection. Section 3 describes the details of various automatic feature selection methods we used. The combination of both data-driven and heuristic domain knowledge-driven methods is presented in Section 4. We describe results from applying automatic feature selection methods and compare the various methods in Section 5 and highlight conclusions from this work.

2. PHYSICS-BASED MODELS

This section describes physics based models. Here we set up the equations of motion for a typical railway vehicle and address the physics-based simulation and data preparation steps and exemplify heuristic feature selection principles.

2.1. Equations of motion

For a better understanding of the system, the dynamic equations of a linear model are set up in three dimensional space. A railway vehicle typically consists of four wheelsets, two bogies (leading and trailing bogie) and a car body as shown in Figure 1. To ensure good ride comfort there are usually two levels of suspension within a railway vehicle. The mechanical coupling elements between wheelsets and bogie such as primary-spring and primary-damper belong to primary suspension level. Secondary suspension level connects the car body to the bogie. The railway vehicle drives along the track with velocity v. Lateral and vertical movements are denoted by y and z. Pitch, roll and yaw angles are denoted by φ, χ and α, respectively (See Figure 1). Knothe and Stichel (2003) derive the formulas of a simple vehicle with two wheelsets, no bogies and a carbody. In the following we will write down the equations of motion for the railway vehicle shown in Figure 1.

Track irregularities in vertical direction, lateral direction, and rolling angle induce oscillations of the bodies. At the first wheelset the input

$$\boldsymbol{u}_1^T = \left[u_{z1}, u_{y1}, u_{\chi1}\right] \tag{1}$$

can be defined as a vector of vertical track irregularities u_{z1}, lateral track irregularities u_{y1} and rolling track irregularities $u_{\chi1}$. The equations of motion can be defined by using Newton´s second law of motion. Car body is indexed by 'CB', bogies by 'BG' (leading bogie: 'BG1', trailing bogie: 'BG2') and wheelsets by 'WS' (e.g. 'WS1' for wheelset 1). The state vector for the car body, can be defined as follows:

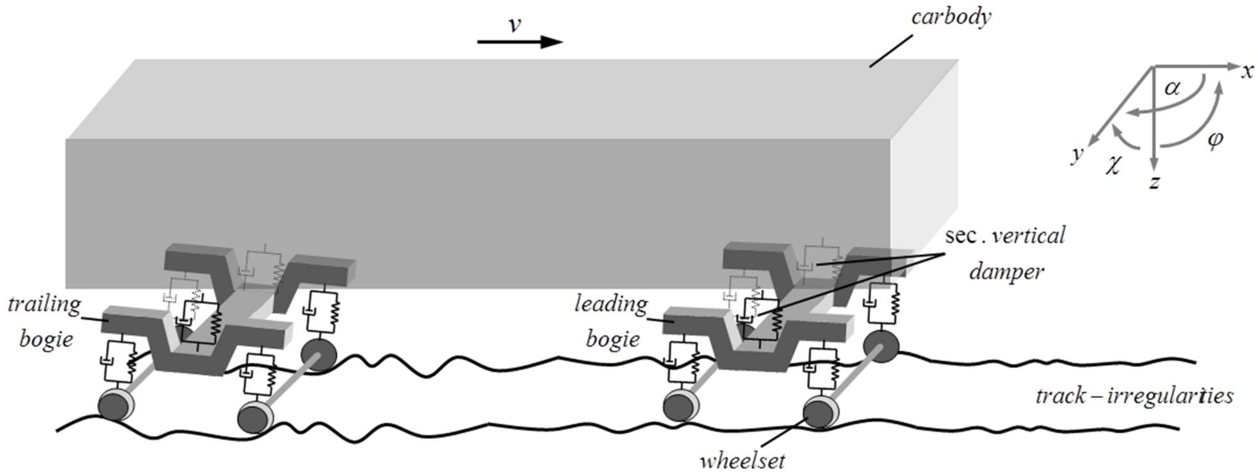

Figure 1. Simplified model of a railway vehicle.

$$\boldsymbol{x}_{CB}^T = [y_{CB}, z_{CB}, \varphi_{CB}, \chi_{CB}, \alpha_{CB}]. \tag{2}$$

The state vectors of the bogie and wheelset can be written in a similar way:

$$\boldsymbol{x}_{BG}^{T} = [y_{BG}, z_{BG}, \varphi_{BG}, \chi_{BG}, \alpha_{BG}] \tag{3}$$

$$\boldsymbol{x}_{WS}^{T} = [y_{WS}, \alpha_{WS}]. \tag{4}$$

The complete system has 23 degrees of freedom (5 for car body, 2x5 for the bogies, and 4x2 for the wheelsets). Combining all state variables into the state vector $\boldsymbol{X}$:

$$\boldsymbol{X}^{T} = [\boldsymbol{x}_{CB}, \boldsymbol{x}_{BG1}, \boldsymbol{x}_{BG2}, \boldsymbol{x}_{WS1}, \boldsymbol{x}_{WS2}, \boldsymbol{x}_{WS3}, \boldsymbol{x}_{WS4}] \tag{5}$$

and all inputs at the four wheelsets into vector $\boldsymbol{U}$:

$$\boldsymbol{U}^{T} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4], \tag{6}$$

we can write the dynamic equations in matrix form as follows:

$$\boldsymbol{M} \cdot \ddot{\boldsymbol{X}} + \boldsymbol{D} \cdot \dot{\boldsymbol{X}} + \boldsymbol{C} \cdot \boldsymbol{X} = \boldsymbol{D}_u \cdot \dot{\boldsymbol{U}} + \boldsymbol{C}_u \cdot \boldsymbol{U}, \tag{7}$$

where $\boldsymbol{M}$ denotes the mass matrix, $\boldsymbol{D}$ and $\boldsymbol{C}$ are the damping and stiffness matrices. $\boldsymbol{D}_u$ and $\boldsymbol{C}_u$ are the damping and stiffness matrices for excitations.

Equation (7) is a multi-dimensional differential equation. Typical numerical methods for integration can handle differential equations of first order. When introducing a state vector:

$$\boldsymbol{Y}^{T} = [\boldsymbol{X}, \dot{\boldsymbol{X}}], \tag{8}$$

the dynamic equations can be written as function of $\boldsymbol{Y}$ and t:

$$\dot{\boldsymbol{Y}} = f(\boldsymbol{Y}, t) \tag{9}$$

Based on (9) numerical methods can be used to solve the equations. Ellermann (2014) describes methods to solve such equations. After solving the equations transfer functions of the system can be calculated. Faulty components induce changes in the transfer functions. The analysis of these changes can be used for feature selection. This will be described in more detail in Section 2.2.

2.2. Data preparation

The relevant failure modes of the system have to be defined. Usually this is done by a Failure Modes, their Effects and Criticality Analysis (FMECA), which defines the critical components and related fault modes (See the International Organization for Standardization [ISO] report, 2012). For example, the fault mode discussed within this work is a faulty secondary vertical damper.

A common challenge in the field of diagnostics is that faulty data is often not available. To tackle this problem healthy and faulty data for relevant fault modes is generated with a virtual railway vehicle model through simulation. A three dimensional non linear model, which can represent the complexities of the mechanical system, was set up in the multibody simulation toolkit Simpack (See Iwnicki, 2006). To handle the various operational conditions like different track layouts and irregularities, payload, speed, Wheel/Rail interaction in combination with different fault modes several methods are defined based on the Design of Experiments (DoE) methodology. The control of the simulator and the experimental conditions ensures that sufficient data is available for machine learning methods. The synthetically generated labeled data is split up into three parts. Training and test data for feature selection methods and validation data to validate models independently.

2.3. Heuristic feature selection

The main idea of the heuristic feature selection is to predefine sensors or signals where changes in the power spectral density or transfer function are visible and develop intuition about the data inputs.

In the examples throughout this paper, the target is to detect faulty secondary damper in the secondary suspension level at the leading bogie. From an engineer's point of view a faulty secondary damper influences the transfer function of the secondary suspension level. After solving the equations described in Section 2.1 the transfer functions can be calculated.

Figure 2 shows the transfer function for vertical acceleration from the leading bogie to the car body. The top subplot shows the transfer function for the nominal (or healthy) state of the car and for a faulty state. An easier comparison of changes in transfer function can be done by referencing the transfer function of a faulty state to a healthy state. The ratio of the transfer functions is plotted on the bottom subplot in Figure 2. Characteristics of the plotted ratio are essential for understanding the effects of a faulty component. Generally a faulty damper leads to higher amplitudes at low frequencies and lower amplitudes at high frequencies.

The dimensionless excitation frequency is defined as

$$\eta = \frac{\Omega}{\omega}, \tag{10}$$

where ω is an eigenfrequency of the system and Ω is the excitation frequency. ω describes the eigenfrequency of an equivalent one mass oscillator. In our case, the system has many eigenfrequencies. For $\eta < \sqrt{2}$ and especially at resonance, a faulty damper leads to higher amplitudes. A faulty damper can result in lower amplitudes, when $\eta > \sqrt{2}$ (See Knothe and Stichel, 2003).

The described behavior is visible in Figure 2. Additionally the roots of the transfer function at certain frequencies are visible. This is an effect of the kinematics of the bogie. The secondary suspension is located in the pitch pole, so that certain frequencies are not transmitted to the car body. These frequencies are influenced by the geometric dimensions such as wheelbase, kingpin distance and velocity.

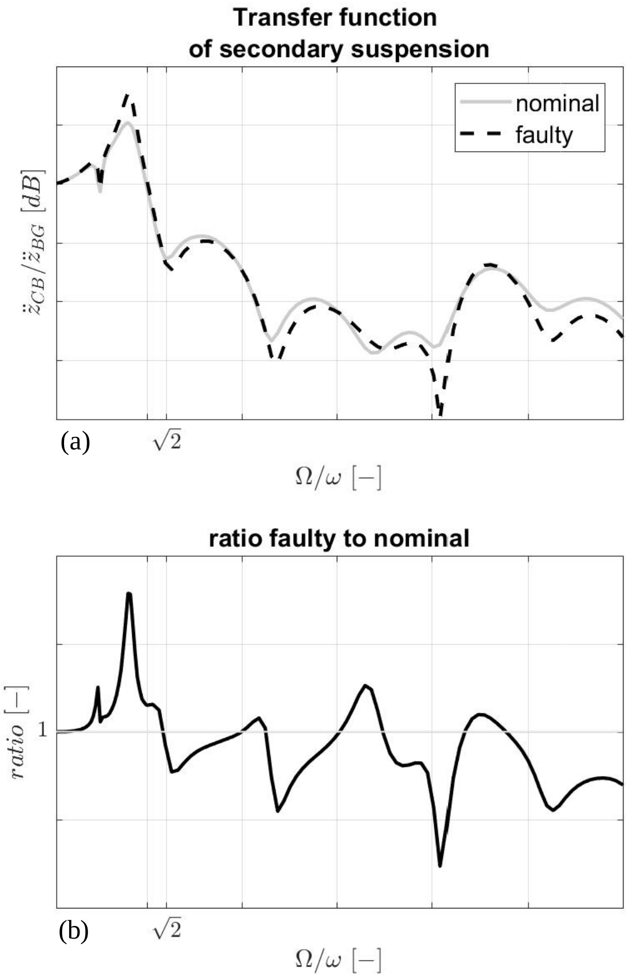

Figure 2. Transfer function $\ddot{\boldsymbol{z}}_{\boldsymbol{CB}}/\ddot{\boldsymbol{z}}_{\boldsymbol{BG1}}$ (a), and ratio faulty to nominal (b).

The influence of velocity on changes in the transfer function is shown in Figure 3. It turns out that possible frequency bands that are good for detection of vertical damper problems strongly depend on velocity. Two good sensors for a classifier of the secondary vertical damper would be acceleration at bogie and car body. The marked frequency bands indicate hypothesized preselection biases for feature selection. These will be checked with data-driven techniques as described further in this paper.

In many practical cases, classification based on a bigger feature subset results in better classification performance because it is easier to find a decision function that separates positive and negative training data. On the other hand larger feature subsets may overcome the risk of “overfitting”. Data overfitting can happen when the number of features is large compared to the number of training examples (See, for example, Guyon, Weston, Barnhill and Vapnik, 2002). Using diverse features in combination is essential for obtaining good classification rates beyond what can be achieved purely based on heuristic methods relying on domain knowledge. For example, the analysis of the power spectral densities of a signal or a transfer function just allows a univariate view. It is very hard to design and visualize multidimensional feature sets. Beside this, there are lots of possible transfer functions within this simplified model (e.g. wheelset to car body, rotational accelerations). An analysis of all of them would be unfeasible.

For a quantitative evaluation of a large, extensive feature set optimization criteria will be defined next.

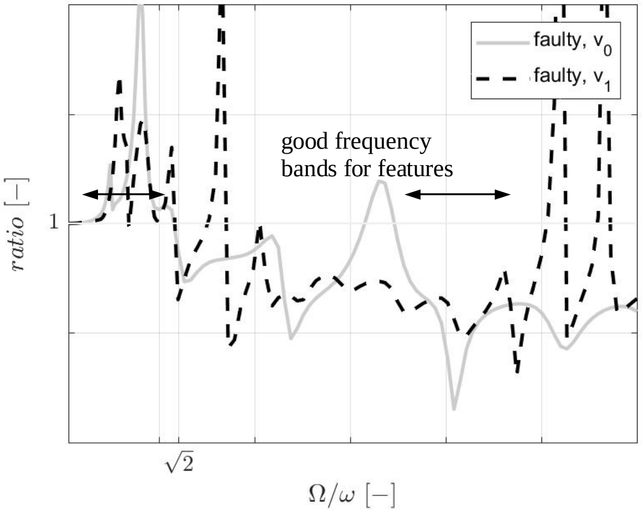

Figure 3. Changes in transfer function at two different velocities indicate what could be good frequency bands for feature definition.

3. AUTOMATED FEATURE SELECTION

Feature selection methods based on data driven techniques can be divided into univariate, multivariate; filter, wrapper and embedded methods (Guyon, Bitter, Ahmed, Brown, and Heller, 2003). Univariate methods consider one feature at a time. In contrast multivariate methods consider subsets of features, which may in some cases result in a better feature set. Filter methods generate a ranking of features without directly optimizing the performance of a classifier, while wrapper and embedded methods are directly linked to a classifier and use misclassification rate to control search in the feature space. Finding the optimal feature subset with wrapper or embedded methods is a rather difficult optimization problem, due to the existence of local minima. In the following, we briefly describe several known methods tested in this work.

3.1. Pairwise Correlations

Pairwise correlations is a filter method. The Pearson correlation coefficient $C_{Pearson}$ captures the linear dependency between random variables x and y. It can be obtained according to the following formula:

$$C_{Pearson}(x,y) = \frac{1}{N-1}\sum_{j=1}^{N}\left(\frac{x_j-\mu_x}{\sigma_x}\right)\left(\frac{y_j-\mu_y}{\sigma_y}\right), \quad (11)$$

where μ_x and σ_x are the mean and standard deviation of x. Each variable has N scalar observations. Good feature sets contain a feature x that is highly correlated with the response variable y. Ranking of features is calculated from the descending order of the $C_{Pearson}(x,y)$. If the correlation coefficient is equal to 1 (or -1), x and y are fully linearly correlated (respectively inversely correlated). A correlation coefficient equal to 0 indicates that there is no linear correlation**.** In that case non-linear correlation is possible (Fisher (1958))**.**

3.2. The Fisher Ratio

An alternative to determine correlation between two time series is Fisher´s ratio coefficient C_{Fisher} . It is calculated by using the mean and standard deviation of labeled data (Fisher (1958)) as follows:

$$C_{Fisher}(j) = \frac{\left(\mu_x-\mu_y\right)^2}{(\sigma_x)^2+\left(\sigma_y\right)^2}. \quad (12)$$

3.3. Kolmogorv-Smirnov Test

Feature selection can be based on the test statistics of a two-sample Kolmogorv-Smirnov test. This nonparametric hypothesis test evaluates if two sample data vectors are from the same continuous distribution. The test statistic is defined as:

$$D^* = \max_x\left(\left|\hat{F}_{1,n}(x)-\hat{F}_{2,n'}(x)\right|\right), \quad (13)$$

where $\hat{F}_{1,n}$and $\hat{F}_{2,n'}$ are the empirical distribution functions of the two samples n and n' (Massey, F.J. (1951)). Again this function is implemented in several applications for data analysis.

3.4. Minimal Redundancy Maximum Relevance

Many used methods do not consider the relationships among features, so that a subset of selected features can be strongly correlated. This method minimizes redundancy and maximizes relevance. Maximum relevance is to search a features set *S* with *m* features {x_i} that satisfy

$$max\, D(S,c),\; D = \frac{1}{|S|}\sum_{x_i e S} I(x_i;c), \quad (14)$$

where *I* denotes the mutual information and *c* the target class. Features selected according to this criterion are correlated, so that the discriminative power does not change when removing a feature that highly depends on another feature. Therefore a minimal redundancy condition can be added to select mutually exclusive features:

$$min\, R(S),\; D = \frac{1}{|S|^2}\sum_{x_i,x_j e S} I\left(x_i,x_j\right). \quad (15)$$

Both constraints get optimized simultaneously by defining a operator $\Phi(D,R)$ (Peng, Long, and Ding, 2005).

3.5. SVM Recursive Feature Elimination

In this method, feature selection is done by a SVM-recursive feature elimination algorithm. We implemented the kernel version of SVM-RFE in (Guyon et al., 2002). This method can equally handle nonlinear SVMs. In order to deal with the problem of many highly correlated features in SVM-RFE, a correlation bias reduction strategy was also part of the algorithm (See Yan and Zhang, 2015).

3.6. Neighborhood Component Analysis (NCA)

NCA is a non-parametric embedded method. The aim is to maximize prediction accuracy of regression or classification algorithms. Our implemented function performs NCA feature selection with regularization. Based on the minimization of an objective function that calculates the average leave one out classification loss, the weights of the features are determined (Yang, Wang and Zuo, 2012).

3.7. Sequential Feature Selection

Sequential feature selection is a widely used method. The method selects a subset of features from a given feature set that best predicts the data by sequentially selecting features until there is no improvement in prediction or the allowed number of features is reached. Selection can be done forward, by adding features when starting from an empty feature subset, or backwards, by removing features when starting with all possible features. It is known that backward selection is computationally more expensive. On the other side forward selection often results in weaker feature subsets because the importance of variables is not assessed in the context of other variables not included yet (Guyon and Elisseeff, 2003). For the selection process, misclassification rate is used as objective function to minimize. There are no restrictions of the classifier type within the search process.

3.8. Alternative Automated Search Algorithm

A new alternative search algorithm, which is based on a combination of univariate ranking, sequential feature selection, and subset combinations is also part of this work. The workflow of the process is shown in Figure 4.

Feature selection starts applying a filter method to find the best univariate features. The best features are used as starting feature for a sequential forward selection, which therefore has to be repeated several times. This results in multiple feature sets, for each of them a model is trained and applied to a test set. All the features of the best models define a new basis feature set. Then all possible combinations with a defined number of features are trained and tested. Using the default values of the parameters the computational costs are acceptable. Compared to backward feature selection, this method is faster.

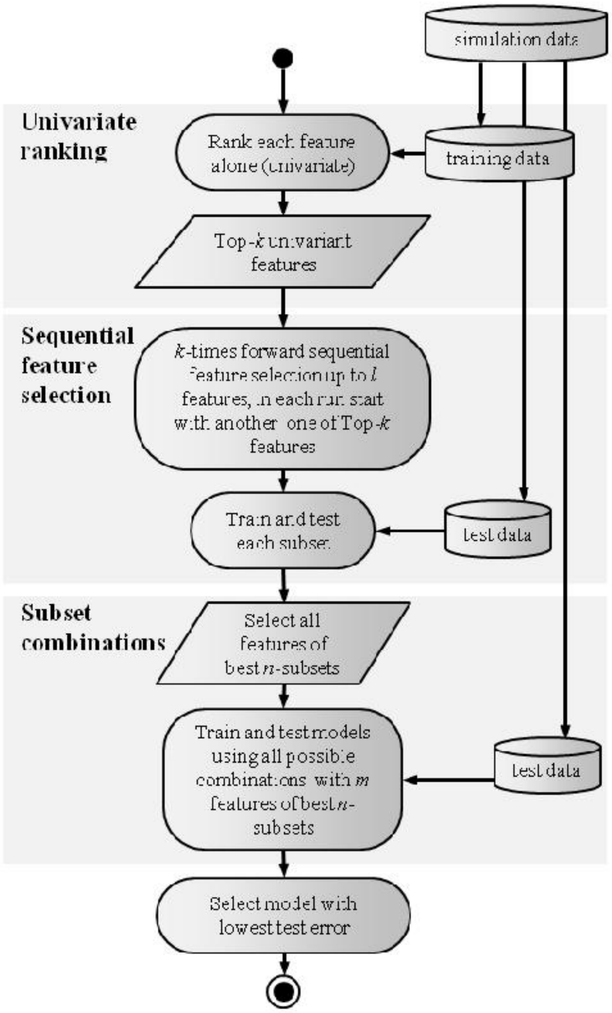
simulation data
Univariate ranking
Rank each feature alone (univariate)
training data
Top-k univariant features
Sequential feature selection
k-times forward sequential feature selection up to l features, in each run start with another one of Top-k features
Train and test each subset
test data
Subset combinations
Select all features of best n-subsets
Train and test models using all possible combinations with m features of best n-subsets
test data
Select model with lowest test error

Figure 4. Workflow of the alternative automated feature selection algorithm.

4. COMBINED FEATURE SELECTION PROCEDURE

Our general workflow for feature selection is shown in Figure 5. The procedure consists of three main tasks. Before starting with the feature selection, data has to be prepared by using physics based models. This includes the definition of relevant failure modes as well as running simulations. After that features get selected by combining heuristic feature selection with automated methods. The usual approach using pure data-driven methods is difficult, as the number of possible feature sets is very large. Based on the 50 sensors that are mounted on the system we can define some thousand different features. The heuristic selection is a rough preselection based on changes in the transfer function, so that the first level subset includes a manageable amount of features. The aim of the heuristic selection is to reduce the feature set for automated feature selection. A smaller feature set produces less computational costs and reduces the risk of finding local minima. Several different methods like filter, wrapper and embedded methods are used within the automated feature selection. The workflow ends with the Validation of selected features and models.

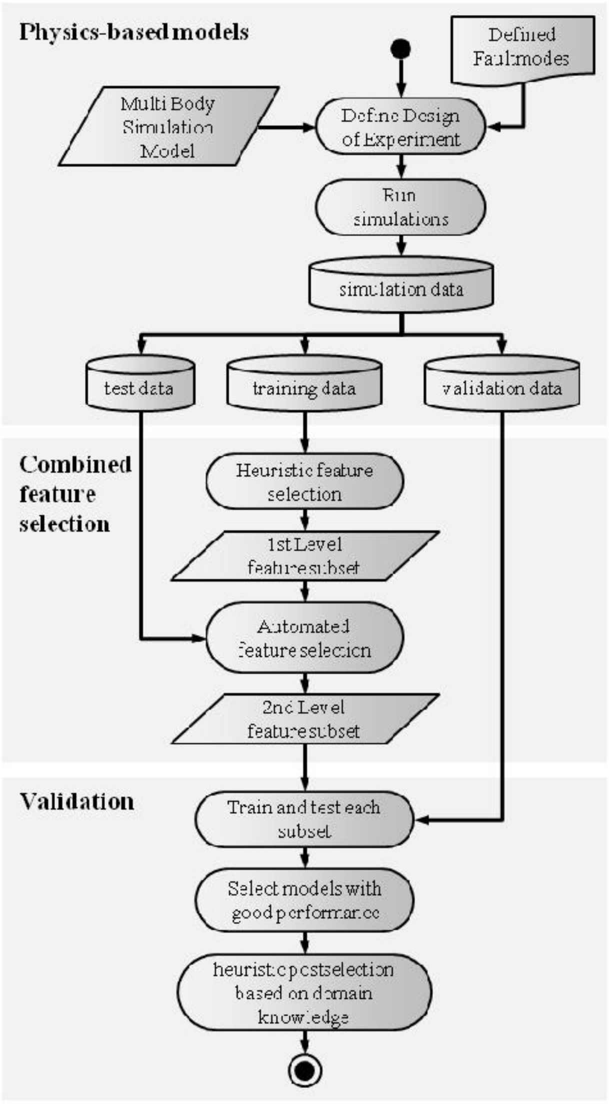
Physics-based models
Defined Faultmodes
Multi Body Simulation Model
Define Design of Experiment
Run simulations
simulation data
test data
training data
validation data
Combined feature selection
Heuristic feature selection
1st Level feature subset
Automated feature selection
2nd Level feature subset
Validation
Train and test each subset
Select models with good performance
heuristic postselection based on domain knowledge

Figure 5. General feature selection procedure

After the feature selection process the validation of the models and feature sets has to be performed. Each feature set gets trained and tested using a classifier. The process ends with post selection based on domain knowledge. At this point we can also take advantage of the symmetry in the design of the railway vehicle. Feature selections for symmetrical fault modes arguably result in symmetrical feature sets. The intensity of a fault mode should only influence the classifier accuracy but should not influence the selected features. Another important point is the size of the selected feature set. A simpler model with just a slightly lower accuracy has always to be preferred, based on generalization arguments.

After ranking the features using heuristic and automated algorithms each feature subset (of up to a acceptable complexity bound, such as ten features in this work) of each method gets trained and tested by using a Support Vector Machine (SVM) classifier (Schölkopf, Williamson, Smola, Shawe-Taylor and Platt, 1999). The criterion for comparing the accuracies of the SVMs used in this work is accuracy, defined as

$$Accuracy = \frac{TP + TN}{ALL}, \qquad (16)$$

where TP is the number of true positives and TN is the number of true negatives. The lower threshold is 0.5.

5. EXPERIMENTAL RESULTS

The results of various automatic feature selection methods are presented in Figure 6. The plot shows the accuracy on validation data when using 1 to 7 features. In case of filter methods the ranked features are used. In the first step feature with rank 1 in the next step feature with rank 1 combined with feature with rank 2 and so on. In case of univariate methods the improvement of accuracy by sequentially adding features is plotted. In general multivariate methods have a much better performance than filter methods. An exception to this is the filter method of Kolmogorv-Smirnov Test, which has an accuracy of 0.85. When comparing the different data-driven feature selection algorithms, forward sequential feature selection gives the best results.

With an accuracy of about 0.93, SVM-recursive feature elimination achieved the second-best result. Pairwise correlation, Fisher ratio and minimal redundancy maximum relevance do not perform very well. They result in an accuracy of approximately 0.7.

Sequential feature selection seems to be the most promising method and will be analyzed in the following in more detail.

Feature selection is a non-convex optimization problem with many local minima. Figure 7, for example, shows the optimization criterion value of sequential feature selection for the first 5 iteration steps. In each step there are a number of minima with often very similar values. The difference between the lowest two minima is indicted in the figure. The minimum of the criterion itself is marked by a square.

As shown, in some steps no distinct minimum of the criterion is visible. Especially at step 2 there are four combinations with almost equal criterion values.

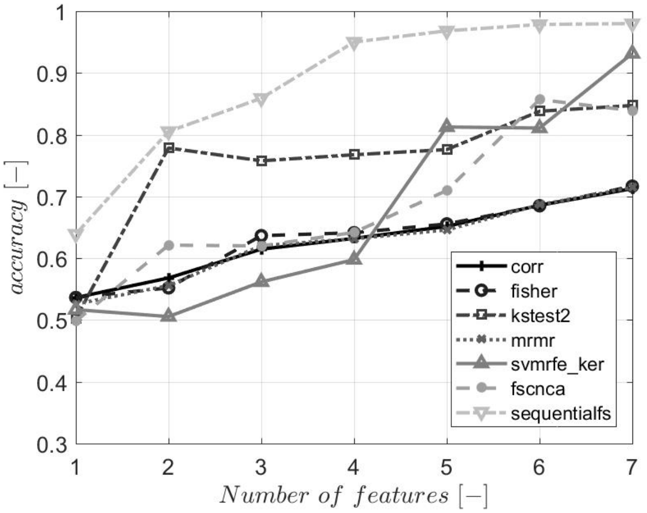

Figure 6. Comparison of performance of automatic feature selection methods.

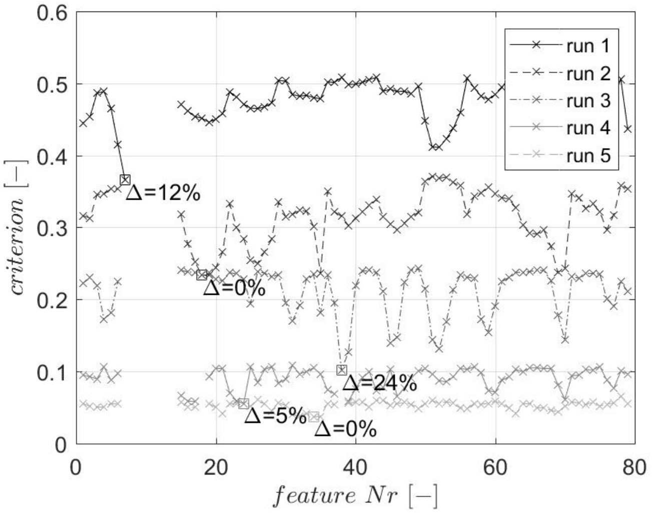

Figure 7. Criterion value of sequential feature selection.

This has two explanations. First, some features are highly correlated because of mechanical coupling. Therefore similar results can be achieved with different combinations of features. Second, in the process multiple random samples are drawn which induces variation to the criterion value. These lead to variation in the results of sequential feature selection when repeating the procedure several times.

Figure 8 shows the results of selection after repeating sequential feature selection 30 times. Each run achieved an accuracy higher than 0.95. The height of the bar for each feature (on the y-axis) illustrates how often the feature was selected. Feature 7 was selected in every single experiment.

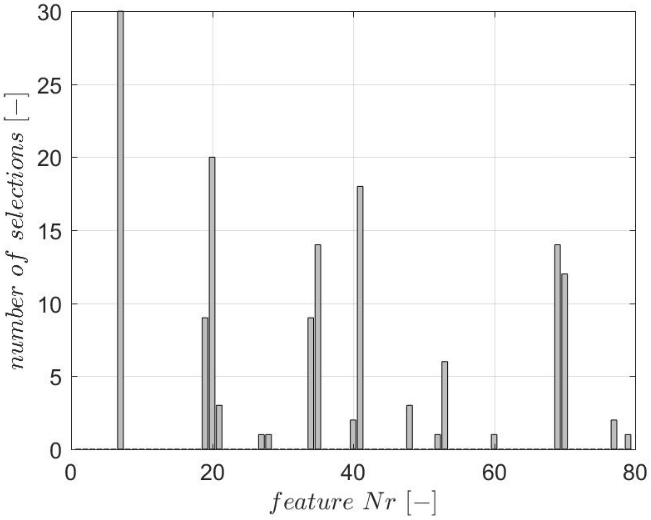

Figure 8. Histogram of the number of selections for a reduced set of features.

This is the feature of the first run when selection is based on a univariate selection. There exist five features that were selected at least ten times.

The results show the following very clearly. Data is noisy but there is sufficient information in it to define a good classifier with a few features. Features can be chosen in many different ways, because there is a lot of redundant information.

Sequential feature selection can be performed either in forward or backward direction. Backward selection is computationally more expensive, but results often in better accuracies (Guyon and Elisseeff, 2003). Figure 9 shows results of forward and backward feature selection starting from approximately 100 features. In our problem backward selection with three features achieves a better accuracy than forward selection with five features. With backward sequential feature selection an additional fourth feature does not really improve accuracy, and a model with a small number of features has always to be preferred. The figure also illustrates that starting with the feature that is univariate the most discriminating one might not always lead to the best result.

As already mentioned backward selection needs more steps during the optimization process than forward selection. Beside the number of models that have to be evaluated, the complexity of the model is also essential for computational costs. During the first steps of backward sequential feature selection the models have to be trained in a high dimensional space, which is computationally expensive.

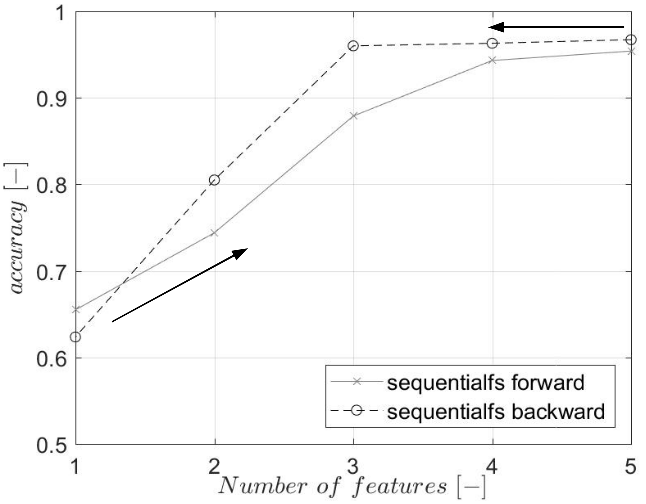

Figure 9. Forward and backward sequential feature selection.

The results obtained so far can be summarized as follows. There is a lot of redundancy in the feature set and there are many different possibilities to define efficient classifiers with a small number of features. Backward selection results in higher accuracy with fewer features, but has high computational costs. Forward and backward sequential feature selection do not fully satisfy our needs and therefore a modified version was designed. One main requirement is that a general view on several feature combinations should be possible.

One advantage of the alternative automated search algorithm is that performing combinations at the end of the workflow will result in several distinct models. Figure 10 shows typical results of our algorithm with default values. The top subplot shows the accuracy of the feature combination. The bottom subplot visualizes all possible subsets of size three or four from a set of seven features. When the feature is marked with a dot in a column it is part of the subset. First half includes combinations with four features; second half includes combinations with three features.

Results show that there are several models with similarly high values of accuracy (greater than 0.95). With four features accuracy is generally higher than with three features. Nevertheless there is also one model with an accuracy higher than 0.95 with just three features. The feature set is similar to the feature set from backward sequential feature selection (See Figure 9).

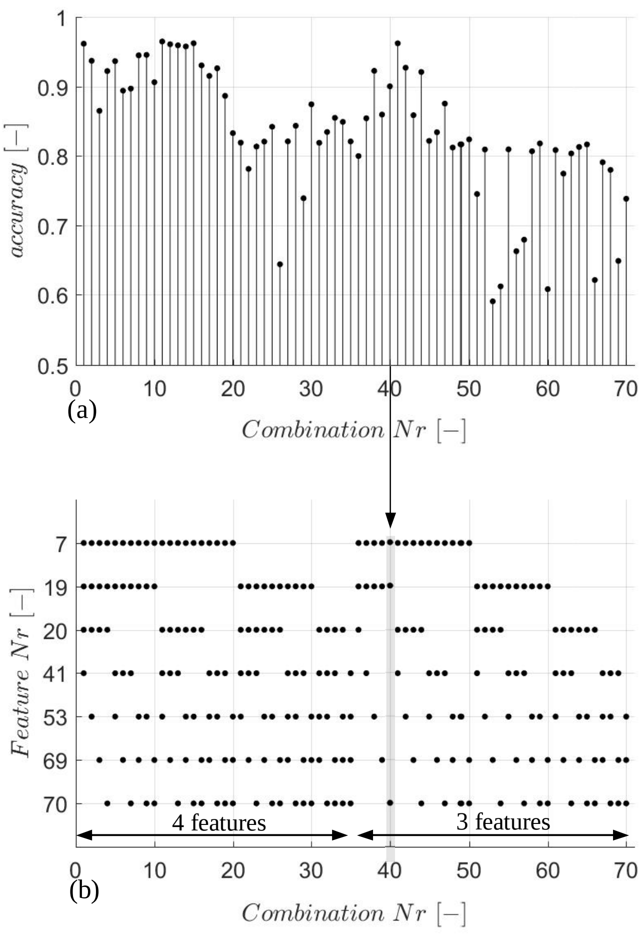

Figure 10. (a) Accuracy of feature combinations; (b) Possible combinations 3 or 4 features.

6. CONCLUSION

The performance of a classifier strongly depends on the input features, which are extracted from raw sensor data. Features should contain as much information as possible regarding a faulty state. A method for feature selection based on domain knowledge in combination with data-driven techniques is presented. The feasibility of the method is shown using simulation data of a railway vehicle. Results show that it is possible to detect a faulty damper with a high accuracy with just a few features. An alternative feature search algorithm, which obtains high accuracies, is also presented. The proposed method was developed for a railway application, but can be used for any other mechanical dynamical system.

REFERENCES

Baumgartner J.P. (2001). Prices and costs in the railway sector, Ecole Polytechnique Federale de Lausanne.

Ellermann (2014). Mehrkörperdynamik, lecture script, Version 18, Graz University of Technology.

Fisher (1958). R.A. Statistical Methods for Research Workers, 13th Ed., Edinburgh : Oliver and Boyd.

Guyon, Bitter, Ahmed, Brown, and Heller (2003). Multivariate Non-Linear Feature Selection with Kernel Multiplicative Updates and Gram-Schmidt Relief proceedings of the BISC FLINT-CIBI 2003 workshop, Berkeley, Dec. 2003.

Guyon I., and Elisseeff A. (2003). An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, pp. 1157-1182.

Guyon I., Weston J., Barnhill St., Vapnik V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning Volume 46, Issue 1, pp. 389-422.

International Standards Organization (ISO) (2012). Condition Monitoring and Diagnostics of Machines - Prognostics part 1: General Guidelines. In ISO, *ISO13379-1:2012(E). vol. ISO/IEC Directives Part 2, I. O. f. S.* (ISO), (p. 2). Genève, Switzerland: International Standards Organization.

Iwnicki S. (2006). Simulation. In Polach O., Berg M. and Iwnicki S. *Handbook of railway vehicle dynamics* (pp. 359-423). Boca Raton London New York.

Kimothol J.K., and Sextro W. (2014). An approach for feature extraction and selection from non-trending data for machinery prognosis, Annual Conference of the Prognostics and Health Management Society. Sept 29 – Oct 02, Fort Worth.

Knothe, K., and Stichel S. (2003). Schienenfahrzeugdynamik. Berlin Heidelberg New York.

Massey, F.J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association, Vol. 46, No. 253, pp. 68-78.

Peng H, Long F., and Ding C. (2005). Feature selection based on mutual in-formation: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238.

Schölkopf B., Williamson R., Smola A., Shawe-Taylor J. and Platt J. (1999). Support vector method for novelty detection. Proceedings of the 12th International Conference on Neural Information Processing Systems pp. 582-588. Nov 29 - Dec 04, 1999 Denver, CO.

Yan K., Zhang D. (2015). Feature selection and analysis on correlated gas sensor data with re-cursive feature elimination, Sensors and Actuators B: Chemical, pp. 353-363.

Yang W., Wang K. and Zuo W. (2012). Neighborhood Component Feature Selection for High-Dimensional Data, Journal of computers, Vol. 7, No. 1.

BIOGRAPHIES

Bernhard Girstmair studied mechanical engineering and business economics at the Graz University of Technology. Since 2012 he is with the Vehicle Dynamics and Analytics group at Siemens Mobility in Graz, Austria. His research focuses on diagnostics and prognostics of mechanical dynamic systems.

Dr. Andreas Haigermoser is principal engineer for bogies at Siemens Mobility in Graz, Austria. He received his doctorate at Graz University of Technology. He has a long experience in bogie engineering and vehicle dynamics. His actual research focuses on diagnostics and prognostics of mechanical dynamic systems.

Dr. Justinian Rosca obtained his Ph.D. in Computer Science from the University of Rochester, NY, and holds also an M.Sc. in Computer Science and an M.Sc. in Computers and Control Engineering. His research interests are in Statistical Signal Processing, Machine Learning, Probabilistic Inference, Artificial Intelligence, and applications in Cyber Physical Systems and Autonomous Systems involving combinations of modeling, simulation, and edge intelligence. Dr. Rosca holds over 50 patents and over 90 publications. He has co-authored two books in mathematics and signal processing. He served as program chair of the 6th Independent Component Analysis and Blind Signal Separation International Conference, chair of the Neural Information and Processing Systems workshop on Sparse Representations in Signal Processing, and recently as chair of the Data Challenge 2015 and 2016 competitions of the Prognostics and Health Management Society.