Incorporating Expert Knowledge into a Self-Organized Approach for Predicting Compressor Faults in a City Bus Fleet

Yuantao Fan, Sławomir Nowaczyk, Thorsteinn Rögnvaldsson
Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Sweden

Abstract.

In the automotive industry, cost effective methods for predictive maintenance are increasingly in demand. The traditional approach for developing diagnostic methods on commercial vehicles is heavily based on knowledge of human experts, and thus it does not scale well to modern vehicles with many components and subsystems.

In previous work we have presented a generic self-organising approach called COSMO that can detect, in an unsupervised manner, many different faults. In a study based on a commercial fleet of 19 buses operating in Kungsbacka, we have been able to predict, for example, fifty percent of the compressors that break down on the road, in many cases weeks before the failure.

In this paper we compare those results with a state of the art approach currently used in the industry, and we investigate how features suggested by experts for detecting compressor failures can be incorporated into the COSMO method. We perform several experiments, using both real and synthetic data, to identify issues that need to be considered to improve the accuracy. The final results show that the COSMO method outperforms the expert method.

Keywords.
Vehicle diagnostics, Predictive maintenance, Fault detection, Receiver Operating Characteristic curve, Expert knowledge

1. Introduction

Unplanned stops are very problematic for commercial heavy vehicles. Component failures on the road often lead to extra damage to other subsystems, extra costs due to towing, failure to meet the transportation schedule and additional waiting time due to unplanned maintenance. One common scenario to prevent such problem is to develop an on-board monitoring method. However, modern vehicles are very complex systems and it is not economically viable to design a specific on-board diagnostic method for every component. The state of the art approaches for developing on-board diagnostic functions rely heavily on knowledge of experienced developers, which takes a long time, many costly experiments and a lot of specification on data analysis. It is not realistic to expect that kind of effort devoted to all faults that can possibly happen on a commercial vehicle. A more automated approach will lead to significant cost savings and self-organizing diagnostic methods that are capable of monitoring various signals and detecting different faults are increasingly demanded in the automotive industry.

A generic approach to improve vehicle uptime was proposed by *Byttner et al.* [1]. It allows for monitoring of on-board sensor streams through the Controller Area Network (CAN) using a special electronic hardware, which is capable of logging various on-board time series and transmitting compressed representations to a computing centre, where those representations are compared across the fleet to find deviations. The general method, called Consensus Self-Organizing Models (COSMO), is based on the idea of "wisdom of the crowd". It assumes that the majority of the vehicles are "healthy" and an individual that deviates from the group should be labelled as potentially "faulty". The deviations are matched against the vehicle service record, which contains the operations performed in the workshop, as well as observations and assessment from the technicians.

In [2] we have proposed a method of evaluating COSMO method on real-world data based on Receiver Operating Characteristic (ROC) curves. The faults we focus on are air system problems, in particular air compressor failures. The air compressor is a vital component: it supplies air in to the gearbox, suspension system, brakes and doors. It is the type of component that rarely breaks and therefore is not included in regular maintenance plan. However, the vehicle will not drive if the air compressor is not working. Our results show that the method can successfully detect half of the failures that occur on the road, with sufficient lead time to schedule repair workshop visits.

In this paper, we have implemented an expert approach, as described in a series of patents by Kenneth A. Fogelstrom [3,4,5], to detect compressor failures using real usage data and compare the performance of this method with COSMO method, a generic self-organized approach. In fact, all current on-board vehicle diagnostic methods are based, to a high degree, on expert knowledge (for some examples see [6,7,8]). It is important that data-driven methods can also take advantage of that existing body of knowledge. Therefore, we have also investigated different ways to incorporate expert knowledge from Fogelstrom's patent into the COSMO method, and evaluate the resulting performance.

2. Related Work

The traditional approach for equipment monitoring and diagnostics on modern transportation systems is usually driven by two ideas: construct a pattern recognition classifier or build a reference model. In both cases, human experts such as workshop mechanics or engineers conduct major part of the development. Some of the technical reviews and state-of-the-art approach can be found in e.g., [9,10,11,12].

In this work we focus on air compressor failures on city buses. The closest work we found related to diagnosing air compressor is a series of US patents [3,4,5], where Fogelstrom presented an approach in detecting problems related to air pressure system on a heavy truck by monitoring the pressure signal from one air tank. The method is based on identifying turning point of air pressure within the tank (e.g. maximum/cut-out and minimum/cut-in pressures, the pressure charge and discharge rate, duration of charging period, etc.). Fogelstrom lists two fault indicators for the air pressure signal: if the pressure never reaches the prescribed maximum, or if the pressure charging time is too long. Likewise, the off-board inspection conducted in workshop for testing the health status of a compressor is to measure the time it takes to reach a certain limit of pressure in the charging test, a description can be found in a troubleshooting guide [13].

3. Method

In this section, we describe the Consensus Self-Organizing (COSMO) method as well as the expert method, based on Fogelstrom's patent, for detecting compressor failures. We propose several ways of utilising expert features to measure the deviation levels of each vehicle, and evaluate both methods using receiver operating characteristic (ROC) curve on real-world data set.

3.1. COSMO

The COSMO method is based on the idea of "wisdom of the crowds" approach and works with a model of the signal. The streaming data is, on daily basis, compressed into a histogram. Those histograms are then cross-compared within the fleet, using Hellinger distance, as a metric, resulting in a pairwise (symmetric) distance matrix **D**.

The z-score for any given vehicle m is computed as follows (for more details see *Fan et al.* [14]). First, the row in **D** with minimum sum is chosen as *most central pattern* (denoted by c) and the z-score is:

$$z(m) = \frac{|\{i = 1, ..., N : d_{i,c} > d_{m,c}\}|}{N}. \tag{1}$$

where $|\cdot|$ denotes cardinality of the set. That is, the z-score for a pattern m is the number of observations that are further away from the most central pattern c than m is. This is, essentially, a non-parametric p-value estimation using the empirical distribution. If m is operating normally, the z-scores should be uniformly distributed between zero and one. The null hypothesis is that all samples are drawn from the same distribution and this hypothesis is tested by comparing the arithmetic mean of the z-score over a certain period (we use 30 days here) with the value expected from a normal distribution [2]. We compute the p-value for $\bar{z}$ using a one-sided test, since we are only interested in the samples that lie at the edge of the distribution, i.e. when the z-score is small.

We denote the deviation level for vehicle v at time t as $x_v(t)$. Each such value indicates its health status and can be used for binary classification (using a warning threshold θ) of the vehicle as either *healthy* or as *faulty*.

3.2. Approach based on expert knowledge

As mentioned in section 2, to detect air compressor failure, Fogelstrom [4] considers the pressure signal within one air tank and proposes several features (e.g. charging rate or maximum pressure) as indicators of a possible problem. He suggests defining a reference model that captures the expected behaviour of the system and measures the deviations between observed features and this reference. For example, the air charging rate is considered as an indicator of how good a compressor's efficiency is. In the patent, he describes a controlled experiment for checking the charging rate of a compressor and how to compare the observed rate with the reference. If the observation is better than the reference, the sample is considered healthy. If the observed rate is below the reference, the sample is considered faulty. The patent also notes that the compressor is driven by the engine, and thus the engine speed should be taken into consideration, but we have found this to only have a minor effect.

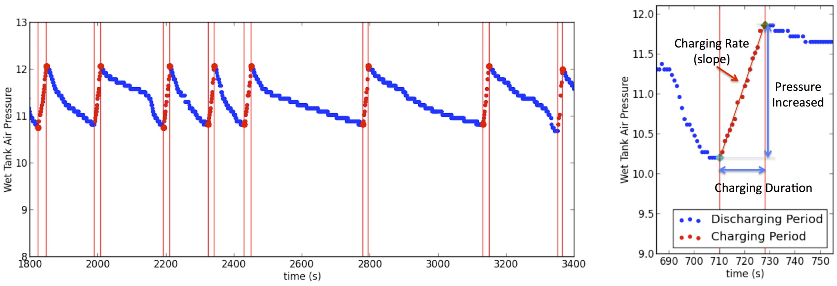

Figure 1. *Wet Tank Air Pressure* signal, with red points corresponding to charging periods and blue points corresponding to discharging periods (left). Expert features that can be extracted from a charging cycle (right).

Figure 1 shows the *Wet Tank Air Pressure* signal of one vehicle during normal operation. As mentioned in Fogelstrom's patent, the pressure signal consists of a charging period (marked with red points) and a discharging period (marked with blue points). The pressure is expected to remain between cut-in and cut-out limits.

We have extracted six features from the *Wet Tank Air Pressure* (shown in figure 1 right): charging rate, maximum and minimum pressure, charging duration, pressure increase and average engine speed during charging period. All six features can be considered as descriptors of the behaviour of the compressor during the charging period and thus should be relevant for detecting the failures. Fogelstrom suggests that different features are useful for different faults. In this work, we evaluate all of them, but we only focus on failures of the compressor itself.

In this work, all features are extracted under vehicle's normal operation. This approach is different from how Fogelstrom have conducted his experiment or how workshop mechanics work. In particular, Fogelstrom built an reference model under controlled experiment and compared the test sample with this reference model. Mechanics perform various tests, e.g. charging time, leakage, etc., within workshop to determine whether the compressor is functioning well enough or needs to be replaced.

The reference model is chosen based on expert's knowledge. However, in this work we evaluate the result using ROC curve and the performance is analysed in a way that is independent from choosing a specific reference model. In particular, for the results presented later, the deviation $x_v(t)$ for vehicle v is computed for each vehicle based on the daily average charging rate.

3.3. Evaluation with Reference Data

The objective of the predictive maintenance is to capture deviations that appear before failures, if any. We assume that the time of the repair, as stated in the vehicle service records, corresponds to the time of component failure. Ideally, deviations should be detected early enough so that actions can be executed to deal with them. For simplicity, we assume here that this period of interest is constant, and refer to it as the *prediction horizon* (PH). For a repair action α performed on vehicle v at time τ, we define the set of *faulty* (or *positive*) samples as:

$$F_v(\alpha) = \{x_v(t) : \tau - \mathrm{PH} \leq t \leq \tau\}. \tag{2}$$

In previous work [2] we have described several types of faults related to air systems and grouped them into five different categories. The expected *healthy* observations are

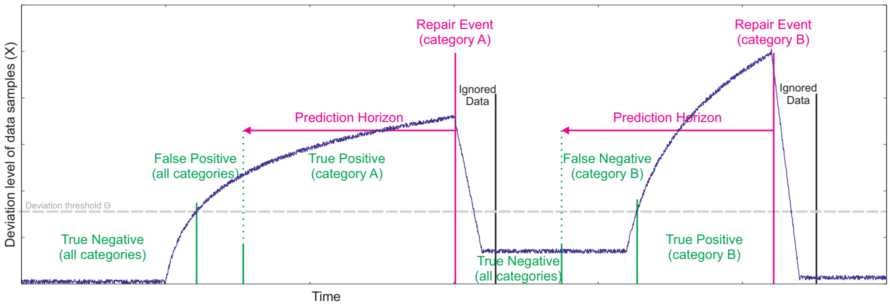

Figure 2. Labelling of deviations in relation to repair actions.

"shared" between all faults, since they correspond to times when a vehicle is believed to be operating without any problem. *Faulty* observations, on the other hand, depend on the particular repair action that was performed. An example of labelling a series of samples is shown in figure 2.

Based on $x_v(t)$ and $F_v(\alpha)$ we can calculate elements in the confusion matrix (true positives, false positives, true negatives and false negatives) for any given threshold θ. By varying θ we can map out the receiver operating characteristic (ROC) curve, which is the relationship between the true positive rate (TPR) and the the false positive rate (FPR). When evaluating the results we use area under the ROC curve (AUC) as the primary quality measure.

4. Experiment Results

Data set Experimental results presented in this paper are based on the data set collected over three years, from June 2011 to September 2014, on a commercial fleet of 19 Volvo buses operated in Sweden. Approximately 100 signals are logged by an on-board embedded device from controller area network (CAN), sampled at 1 Hz. Date and type of repair actions were collected from manually curated vehicle service records.

Deviation level labelling Figure 3 shows the labelled deviation level for bus *A* using COSMO method (top) and expert method (bottom). Vertical lines correspond to repairs performed: red is compressor replacement that required towing, while blue indicate other

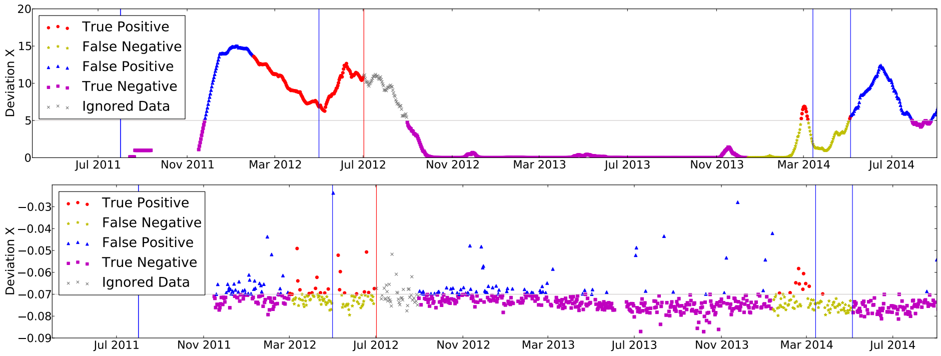

Figure 3. Deviation level of bus *A* using COSMO method (top) and expert method (bottom). Vertical lines correspond to repairs performed: red is compressor replacement that required towing, while blue indicate other faults related to the air system. Gray lines are correspondent to threshold θ.

faults related to the air system. True positive, false negative, false positive and true negative samples are illustrated using different colors. We use a prediction horizon of 60 days since it is enough to schedule and perform repair. The two month period after a compressor was replaced is excluded since a new compressor often performs very differently from regular ones.

It is important to observe that COSMO produces a measure that aggregates the estimation over 30 day periods, and thus the deviation level builds up incrementally and falls off smoothly. On the other hand, the deviation level of expert method is computed by comparing observed features with the reference model on a daily basis, without any such aggregation.

ROC curve analysis As can be seen in Figure 4, the best performance is achieved by the COSMO method when vehicle comparisons are based on daily histograms calculated only from *Wet Tank Air Pressure* values that belong to charging periods (AUC is 0.75 ± 0.16). COSMO method using all *Wet Tank Air Pressure* values worked slightly worse, with AUC of 0.72 ± 0.16. Next was the expert method based on average values of the charging rate (AUC equal to 0.66 ± 0.04). Finally, the COSMO method using daily histograms of charging rates performed the worst, with AUC of 0.53 ± 0.16.

It is remarkable that a generic unsupervised method like COSMO performs better than an expert knowledge driven approach. The expert method requires a lot of manual work in analyzing the signal, defining features and building a reference model for each specific problem. In contrast, the COSMO method has a simple setup, cross comparing the models computed from raw signals, and thus can be applied on other signals without putting effort into modifying the algorithm. When COSMO method successfully detects a component which will break on road, extra costs due to unplanned stop are saved. *Prytz at el.* [15] have proposed that a planned compressor replacement is ten times cheaper than an unplanned maintenance caused by compressor road side breakdown. Under this circumstance, COSMO method is cost effective.

We initially expected that the COSMO method would achieve better performance after incorporating expert features, rather than only using histograms computed from raw pressure readings, and that the magenta plot on Figure 4 would be the best. Instead, it turns out that using the charging rate, the descriptor that is supposed to capture the

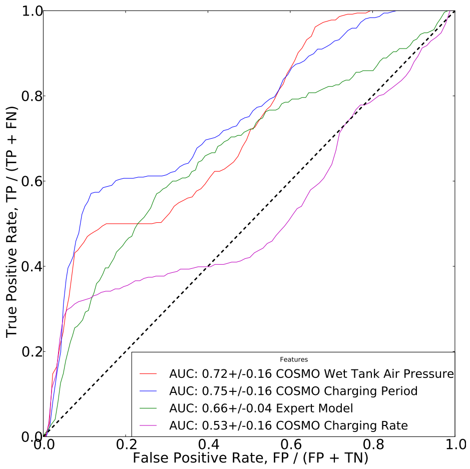

Figure 4. Comparison of ROC curves for predicting compressor failures: the COSMO method using three different features, as well as the expert method using charging rate.

primary interesting feature of the charging period works approximately as bad as random guessing. However, what has proven useful was limiting the raw sensor readings to only the pressure values within the charging periods, with the pressure readings from discharging period removed.

In addition, we analysed the effect of incorporating different expert features into COSMO method. As shown in figure 5, expert method by using charging rate achieves the best performance with 0.66 ± 0.04, using pressure increased is slightly worse, with an AUC of 0.63 ± 0.04, while using maximum pressure achieves 0.59 ± 0.04, AUC of using minimum is 0.48 ± 0.04, using engine speed is 0.46 ± 0.04 and using charging duration is 0.56 ± 0.04. AUC of COSMO method using maximum pressure is slightly worse than expert method using charging rate, within AUC of 0.63 ± 0.16, while AUC of using engine speed is 0.61 ± 0.16, followed by using charging duration, 0.59 ± 0.16, using minimum pressure is 0.56 ± 0.16, using charging rate is 0.53 ± 0.16 and using pressure increased is 0.52 ± 0.16.

Synthetic Experiment One of the major distinctions between the two method is that they are using different models to represent the data: the COSMO method uses histograms and expert method uses averages. In the following experiment we use synthetic data to investigate the effect of using histogram and average value for distinguishing between two signals with different distributions.

Conceptually, we introduce two different distributions representing to three groups of vehicles: the *reference* group, the *healthy* group, and the *faulty* group. For each vehicle we generate data and calculate appropriate model. We then compare the models from healthy and faulty groups against the reference, classify them into two categories and evaluate the result using AUC. Any vehicle i is represented by a parameter pair drawn as: $\mu_k^i \sim N(M_{err}^k, 1)$ and $\sigma_k^i \sim N(V_{err}^k, 0.2)$, where M_{err}^k and V_{err}^k is the meta difference (in the mean and variance) between healthy and faulty groups, at level k. Parameter pair (μ_0, σ_0) corresponds to *healthy* or *reference* vehicles, while (μ_k, σ_k) parameter pairs (with non-zero k) correspond to *faulty* vehicles. For any vehicle i, we draw 500 samples as observations: $Sample^i \sim N\left(\mu^i, \sigma^i\right)$. Each of the *healthy*, *reference* and, for a given k, *faulty* groups consists of 100 vehicles. From the generated data, histograms and average values are computed as models, and distances between those models are calculated, comparing *healthy* and *faulty* vehicles against the *reference* groups.

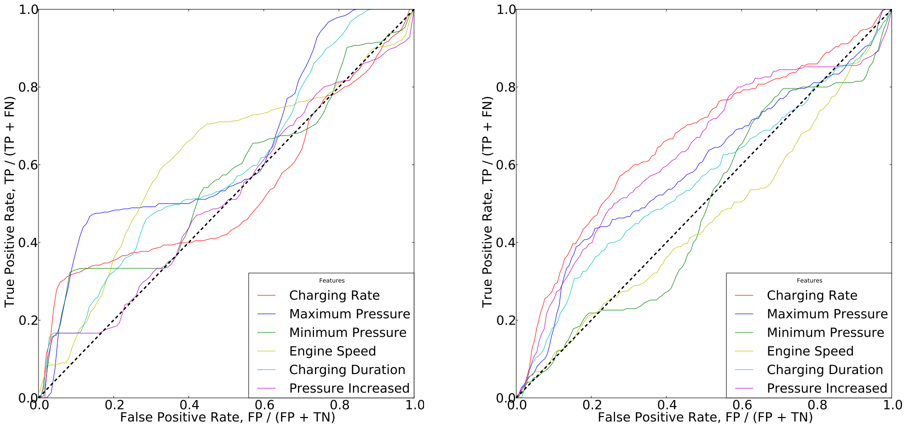

Figure 5. ROC curve comparison by using different method, using the COSMO method (left), using expert approach (right).

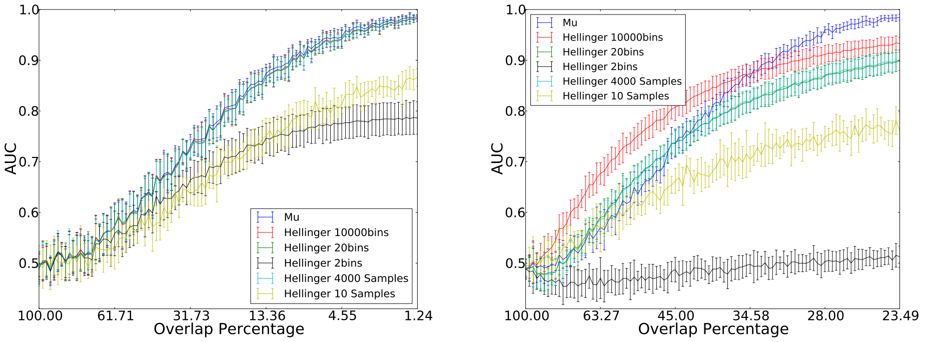

Figure 6. AUC of comparing two set of samples using different method, only M_{err} is varied (left), both M_{err} and V_{err} are varied (right).

Figure 6 presents the effect of different models on classifier performance, as the two distributions (healthy and faulty vehicles) become less similar. This is expressed in terms of overlapping area between the two distributions (parameter pairs). Figure 6 (left) shows the area under curve in terms of overlapping area when only the meta difference M_{err}^{k} is varied. It can be seen that using Hellinger distance between histograms achieves the same performance as using difference between means, except for the cases when very low number of bins (illustrated using yellow curve) or insufficient number of samples (shown in black) are used. Figure 6 (right) shows the AUC by varying both the M_{err} and V_{err}. Measure based on $\Delta\mu$ outperforms Hellinger distances only if the two distributions have an overlapping area of less than 30%. For more diverse distributions, $\Delta\mu$ and Hellinger distances achieve similar results. Thus, the choice of which of the two dissimilarity measures should be used depends on difference between the healthy and faulty distributions.

Analysis of histogram models Figure $7a$ and $7b$ shows the distributions of charging rates from positive (*faulty*) samples, in red histogram, and negative (*healthy*) samples, in green histogram. The two histograms in figure $7a$ are computed based on individual charging rates from two classes, while the ones in figure $7b$ use average values of charging rates over one day period. The overlapping area between the two distributions directly influences the performance of classifier trying to distinguish between the two sets of samples. If most of the samples from the two classes are merged together, it is not possible to distinguish between them without introducing additional features. In figure $7a$, the overlap of the two histograms is 0.92 and in $7b$ it is 0.58. This means that using daily average of charging rates should lead to better classification accuracy than using all charging rates individually.

However, it is clear that difference in a single statistical moment like average value is not sufficient for classification. As a way to introduce more dimensions for comparing different samples, we decide to use histograms, as approximations of the density functions. We compute Hellinger distances between histograms of two negative samples, denoted as $dist(neg,neg)$, as well as distances between histograms of positive samples and negative samples, denoted as $dist(pos,neg)$. The two distributions of distances are shown in figure $7c$ and $7d$. The overlap of two histograms when using Hellinger distance, shown in figure $7c$, is 0.9 and the overlap using $\Delta\mu$, in figure $7d$, is 0.84. The AUC of ROC curve by Hellinger distance in this case is 0.55 ± 0.004 while AUC of using $\Delta\mu$ is 0.62 ± 0.008. This result shows that differences in average charging rate is more discriminative than Hellinger distances. This is consistent with the ROC curve shown in figure 4.

As a final experiment, we investigate the effect of using either pressure values or their aggregated descriptors (i.e. charging rates) for distinguishing between two sets of samples. As before, we introduce two different distributions to represent two groups of vehicles: the *healthy* group and the *faulty* group. For each vehicle we generate data samples, calculate corresponding charging rates, and compute the model (a histogram of either the pressure values or the charging rates). We calculate Hellinger distances between the models and compute the overlap between the distributions, shown on Figure 7: *faulty* against *reference* samples in red histogram and *healthy* against *reference* samples in green histogram. In particular, figure $7e$ and $7g$ show the overlap of using models based on charging rates, while figure $7f$ and $7h$ show the overlap based on data samples.

For generating all *healthy* samples we use variance σ. On the other hand, for creating *faulty* samples in figure $7e$ and $7f$ we use $\sigma + \Delta\sigma_{var}$ is employed in, while for figure $7g$ and $7h$ we use $\sigma + 2\Delta\sigma_{var}$. The overlap in figure $7e$ is 0.33, in figure $7f$ it is 0.56, in figure $7g$ it is 0.32 and, finally, in figure $7h$ it is 0.06. It can be observed that the overlap of using sample descriptor is smaller than using samples i in cases where a small difference in variance is introduced in *faulty* samples (figure $7e$ and $7f$). However, if a larger variance difference is introduced, the overlap by using samples is smaller than by using sample descriptors. The classification performance of using samples and sample descriptor depend on how large the differences is in the variance of *healthy* samples and *faulty* samples. Therefore, one possible reason for the weak performance the COSMO method using charging rates is that there are large difference in variance between *healthy* and *faulty* samples.

5. Conclusions

In this papers we have shown that an generic unsupervised deviation detection method (COSMO), based purely on monitoring available signal and comparing it against other vehicles in the fleet, can detect compressor failures more accurately than an expert

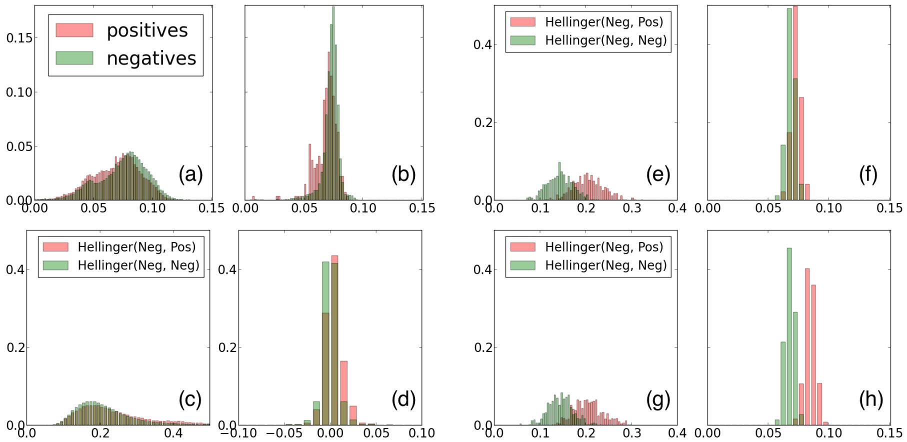

Figure 7. Distribution of charging rates belonging to the two class, individual charging rates (a), daily average of charging rates (b). Distributions of the two distance measures between different sample sets randomly draw from two classes, Hellinger distances (c), difference in μ (d). Overlaps of the two distances distributions using different feature inputs: (e) and (g) are computed from histograms of descriptors, (f) and (h) are computed from histograms of pressure values. The pressure values generated in (g) and (h) have a larger variance compare to (e) and (f).

knowledge driven approach. This is an important results from practical point of view, since in the automotive industry cost effective methods for predictive maintenance are increasingly in demand.

The main contribution of this paper is the comparison between different variations of COSMO methods and an approach based on expert knowledge. In addition, we have implemented the expert method on our data set and analyzed the effect of incorporating the existing expert knowledge into the self-organizing approach. We evaluate several features that experts suggest are useful for detecting compressor failures, as described in a recent patent. The performance of the COSMO method is evaluated when using both the raw data as well as those features. We perform several experiments, using both real and synthetic data, to explain some of the non-intuitive observations. The final results show that the COSMO method augmented with expert features outperforms other solutions.

References

[1] S. Byttner, T. Rögnvaldsson, M. Svensson, Consensus self-organized models for fault detection (COSMO), Engineering Applications of Artificial Intelligence 24 (2011) 833–839.

[2] Y. Fan, S. Nowaczyk, T. Rögnvaldsson, Evaluation of self-organized approach for predicting compressor faults in a city bus fleet, Procedia Computer Science 53 (2015) 447–456.

[3] K. A. Fogelstrom, Air brake system characterization by self learning algorithm (2006 (filed 2004)).

[4] K. A. Fogelstrom, Prognostic and diagnostic system for air brakes (2007 (filed 2005)).

[5] K. A. Fogelstrom, Air brake system monitoring for pre-trip inspection (2008 (filed 2004)).

[6] S. A. Mostafa, M. S. Ahmad, M. A. Mohammed, O. I. Obaid, Implementing an expert diagnostic assistance system for car failure and malfunction, IJCSI International Journal of Computer Science Issues 9 (2) (2012) 1694–0814.

[7] A. T. Al-Taani, An expert system for car failure diagnosis., IEC (Prague) 5 (2005) 457–560.

[8] M. Ruta, F. Scioscia, F. Gramegna, E. Di Sciascio, A mobile knowledge-based system for on-board diagnostics and car driving assistance, in: UBICOMM 2010, The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Citeseer, 2010, pp. 91–96.

[9] R. Isermann, Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance, Springer-Verlag, Heidelberg, 2006.

[10] J. Hines, R. Seibert, Technical review of on-line monitoring techniques for performance assessment. volume 1: State-of-the-art, Technical review NUREG/CR-6895, U.S. Nuclear Regulatory Commission, Washington, DC 20555-0001 (2006).

[11] J. Hines, D. Garvey, R. Seibert, , A. Usynin, Technical review of on-line monitoring techniques for performance assessment. volume 2: Theoretical issues, Technical review NUREG/CR-6895, Vol. 2, U.S. Nuclear Regulatory Commission, Washington, DC 20555-0001 (2008).

[12] J. Hines, J. Garvey, D. R. Garvey, R. Seibert, Technical review of on-line monitoring techniques for performance assessment. volume 3: Limiting case studies, Technical review NUREG/CR-6895, Vol. 3, U.S. Nuclear Regulatory Commission, Washington, DC 20555-0001 (2008).

[13] Bendix Commercial Vehicle Systems LLC, Advanced Troubleshooting Guide for Air Brake Compressors (2004).

[14] T. Rögnvaldsson, H. Norrman, S. Byttner, E. Järpe, Estimating p-values for deviation detection, in: 8^{th} IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO 2014), London, UK, September 8-12, 2014, IEEE Computer Society, 2015, pp. 1–4.

[15] R. Prytz, S. Nowaczyk, T. Rögnvaldsson, S. Byttner, Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data, Engineering applications of artificial intelligence 41 (2015) 139–150.