Citation for the original published paper:

Nowaczyk, S., Prytz, R., Rögnvaldsson, T., Byttner, S. (2013)
Towards a Machine Learning Algorithm for Predicting Truck Compressor Failures Using Logged Vehicle Data.
In: Manfred Jaeger, Thomas Dyhre Nielsen, Paolo Viappiani (ed.), *Twelfth Scandinavian Conference on Artificial Intelligence* (pp. 205-214). Amsterdam: IOS Press
Frontiers in Artificial Intelligence and Applications
http://dx.doi.org/10.3233/978-1-61499-330-8-205

Towards a Machine Learning Algorithm for Predicting Truck Compressor Failures Using Logged Vehicle Data

Sławomir Nowaczyk [a] Rune Prytz [b] Thorsteinn Rögnvaldsson [a] Stefan Byttner [a]

[a] *Center for Applied Intelligent Systems Research, Halmstad University, Sweden*

[b] *Volvo Group Trucks Technology, Advanced Technology & Research, Göteborg*

Abstract. Predictive maintenance is becoming more and more important for the commercial vehicle manufactures, as focus shifts from product- to service-based operation. The idea is to provide a dynamic maintenance schedule, fulfilling specific needs of individual vehicles. Luckily, the same shift of focus, as well as technological advancements in the telecommunication area, make long-term data collection more widespread, delivering the necessary data.

We have found, however, that the standard attribute-value knowledge representation is not rich enough to capture important dependencies in this domain. Therefore, we are proposing a new rule induction algorithm, inspired by Michalski's classical AQ approach. Our method is aware that data concerning each vehicle consists of time-ordered sequences of readouts. When evaluating candidate rules, it takes into account the composite performance for each truck, instead of considering individual readouts in separation. This allows us more flexibility, in particular in defining desired prediction horizon in a fuzzy, instead of crisp, manner.

Keywords. Machine Learning, Relational Learning, AQ, Fault Prediction, Automotive Diagnostics, Logged Vehicle Data.

1. Introduction

Commercial vehicle manufactures are becoming increasingly interested in predicting the current maintenance needs of individual vehicles. This is due to the trend in the sector where OEMs, instead of providing trucks, are heading towards selling transport services. This means moving the operational risk from the customer to the manufacturer. A typical transport service product could be a fleet of trucks, and it would often include maintenance and service plan that guarantees a certain level of availability.

In those settings, statistical lifetime predictions are no longer sufficient. Instead, there is a need for new data mining and machine learning algorithms. The time to market for new data-driven predictive maintenance functionality is long, mostly due to the time it takes to log and process the data. In many cases, however, this can be shortened by taking advantage of the information that is already available in existing data sources. An example would be *Logged Vehicle Data* (LVD), *Vehicle Service Records* and *Vehicle Data Administration* used by Volvo. Those databases have been designed, and are used daily, for other purposes. They are, therefore, already filled with relevant data.

In this paper we make use of all three of those databases, but the most important one for us is LVD, which contains aggregate information about vehicle usage patterns. During normal truck operation, a number of such parameters are continuously collected, and they are downloaded each time a truck is serviced at a Volvo Authorised Workshop. This generates a new data point several times per year, however, at intervals that are unknown *a priori* and difficult to predict even during operation. Nevertheless, a system that continuously monitors incoming data may look for signs of imminent component failure and flag suspicious vehicles to the workshop personnel, who would then schedule replacements, possibly after performing additional tests.

In our previous work [12] we have applied three traditional machine learning algorithms (Random Forest, C5.0 and KNN) to the problem of predicting compressor failures based on the LVD data. However, the results, while encouraging, were not entirely satisfactory. One of the limiting factors was insufficient expressiveness of the classical attribute-value knowledge representation. It is easy to represent individual data readout in this form, but there is no way to take into account the dependencies existing between subsequent readouts from the same truck.

In particular, defining positive and negative training examples poses a problem. The natural way is to specify a *prediction horizon*, and to define that readouts shortly followed by compressor failure should be treated as positive examples, while those that occurred much earlier should be negatives. However, from a business perspective, boundary between those two cases is definitely not crisp.

Therefore, in this paper we report our experiences in designing a Rule-based classification algorithm with Relaxed Prediction horizon (RRP). This algorithm allows the boundary between faulty and non-faulty state to be automatically determined during learning, individually for each vehicle. This allows us to capture the important aspect of maintenance in the automotive industry: events and changes in usage patterns that precede actual failures vary significantly, but can greatly increase component wear.

This paper is organised as follows. In the next section we describe in more detail the type of data we are working with, as well as present the business constraints that dictate how we state the problem and how are we trying to solve it. We follow with a discussion of related research in Section 3. In Section 4 we present our approach, and we outline the algorithm we are working on in Section 5. Results of experiments we have conducted are described in Section 6. We close with conclusions and plans for future work.

2. Databases

Logged Vehicle Data is a Volvo internal database which gathers usage and ambient statistics from Volvo vehicles. The data is downloaded from the truck when it is serviced at an authorised workshop, or wirelessly through a telematics gateway. The database is used during product development, after market and even for sales support.

This database contains data of varying types and has high number of dimensions. Typically a vehicle record contains hundreds of parameters and at most tens of readouts. The number of readouts depends on the availability of telematics equipment and on whether the vehicle has been regularly maintained at a Volvo workshops. For example, in our dataset the average number of readouts is 4 per vehicle per year. However, the variance is very high and many trucks have no readouts at all.

The Volvo Service Records (VSR) is a database that keeps track of all maintenance and repair operations done on a particular vehicle. The database is mainly used by the workshop personnel for invoicing purposes, as well as for diagnostics, allowing access to history of previously carried out repairs.

A typical event contains date, current mileage, and a list of unique maintenance operation codes and exchanged part numbers. In addition to that there may be a text note added by a technician. For the purposes of this work, we are using VSR to find out whether and when a compressor was replaced on a given truck. Our data consists of close to 4000 trucks, out of which 180, i.e. approximately 5%, have had compressor replaced.

3. Related Work

In a survey of Artificial Intelligence solutions being used within automotive industry, Gusikhinet et al.[6] discuss, among other things, both fault prognostics and after-sales service and warranty claims. In a representative example of work being done in this area Buddhakulsomsiri and Zakarian[1] present data mining algorithms that extract associative and sequential patterns from a large automotive warranty database, capturing relationships among occurrences of warranty claims over time. In that work, however, no information about vehicle usage is available, and the discovered knowledge is of a statistical nature concerning relations between common faults, rather than describing a concrete individual.

More recently Choudhary el al.[3] presented a survey of 150 papers related to the use of data mining in manufacturing. While their scope was broader than only diagnostics and fault prediction, including areas such as design, supply chain and customer relations, they have covered a large portion of literature related to the topic of this paper. The general conclusion is that the specifics of automotive industry make fault prediction a more challenging problem than in other domains: almost all prior research considers a case where continuous monitoring of devices is possible.

It is also becoming more common to consider emergent solutions, where vehicles are able to communicate using telematic gateways. Kargupta et al.[7] in an early paper show a system architecture for distributed data-mining in vehicles, and discusses the challenges in automating vehicle data analysis. Zhang et al.[14] show that cross-fleet analysis, i.e. comparing properties of different vehicles, benefits root-cause analysis for pre-production diagnostics. Byttner et al.[2] propose a method called COSMO for distributed search of "interesting relations" among on-board signals in a fleet of vehicles, enabling deviation detection in specific components.

A method based on a similar concept of monitoring correlations, but for a single vehicle instead of a fleet, is shown by D'Silva in [4]. Vachkov in [13] uses a neural gas algorithm to model interesting relations for diagnostic of hydraulic excavators. Contrary to our work, however, both the papers by D'Silva and Vachkov assume that signals which contain the interesting relations are known *a priori*. Lacaille and Come in [9] present a method for monitoring relations between signals in aircraft engines. Relations are compared across a fleet of planes and flights. Unlike us, however, they focus on discovering relationships that are later evaluated by domain experts.

The work presented in this paper falls within the area of Relational Learning or Relational Data Mining. For an introduction to the field, see for example [5] or [8]. Since the early work of Muggleton on Inductive Logic Programming in [11], researchers have been interested in exploring more expressive knowledge representations in conjunction with machine learning. In practice, however, a lot of those methods have had hard time getting industrial acceptance. Therefore, it is important to look for good matches between algorithms and real-world problems. We believe this to be one of them.

4. Approach

In this work we have decided to focus on analysing the air compressor. It supplies the brakes, gearbox, suspension and auxiliary loads with power for operation. A sudden loss of compressed air supply results in locked brakes and immediately halts the vehicle. Such unplanned stops are very costly, both for the customer and for the OEM. For trucks operating under service contract or within warranty, expenses such as towing, disruption of garage workflow, actual repair, rent of replacement truck and loss of OEM's reputation would typically be incurred.

A significant portion of those costs could be avoided by using predictive maintenance, but obviously there is also potential for added costs. Those typically come from unnecessary repairs performed on the incorrectly diagnosed vehicles, as well as from wasted component life. The importance of the latter factor varies greatly depending on particular application. In our case, since components such as compressor or turbocharger are exchanged at most once during a vehicle's lifetime, it can be mostly ignored. Therefore, we mostly focus on unnecessary repairs.

In an earlier work [12] we have considered each LVD readout as a single learning example. This gave us flexibility in both the classifier choice and in deciding how to analyse actual faults. In the following sections we will restate some of those results. At the same time, the new work presented in this paper is motivated by the shortcomings we have identified during that analysis. We believe that by assuming independence between readouts we are losing a very important piece of information.

We have defined *Prediction Horizon* (PH) as a fixed time interval, and classifying readouts as either "less than PH until failure" or "more than PH until failure". This approach works fine in many domains, but is not well suited for our case. The main problems are the unpredictability of intervals between data readouts and the variability in usage patterns of vehicles. We believe now that there is no universal value of prediction horizon that would be suitable for all trucks.

As an evaluation of the business case for the predictive maintenance solution, we introduce measure of cost savings:

$$C_{save} = TP \cdot (C_u - C_s) - FP \cdot C_p$$

The predictive maintenance method will be profitable if the correctly classified faulty trucks (i.e. *true positives, TP*) save more money than the non-faulty trucks wrongly classified as faulty (i.e. *false positive, FP*) waste. There is a delicate balance to be found. On the one hand, the cost of an on-road *unplanned* breakdown (C_u) is much higher than the cost of *scheduled* component replacement (C_s), however, the number of trucks without faults is much higher than the number of trucks with faults.

5. RRP Algorithm

The main contribution of this paper is documenting our work on developing a specialised rule induction algorithm that is capable of working with the complete data concerning a particular truck, and does not assume independence of individual LVD readouts. We have decided to use a bottom-up approach, inspired by the classical family of AQ algorithms by Michalski [10]. The basic idea is to perform iterative rule induction: in each step, we start by selecting a seed example and generate the “best” rule that covers it. This process is repeated until we run out of seed candidates.

Over the years many researchers have proposed various extensions to this core process, but at this stage, we are mainly interested in figuring out whether the general idea is a feasible approach for this particular problem. Therefore, we keep the algorithm as simple as possible. We have identified a number of directions to investigate in the Future Work section.

The RRP algorithm works as follows. Initially, the set of rules is empty, so any example can be used as the seed. In our case, we select a random truck from among those that have had compressor failure. The next step is to generate a rule that covers this example, and as many other positive examples as possible, without covering too many negative ones. As we mentioned earlier, however, the notion of positive and negative examples is somewhat fuzzy in our setting.

Therefore, we start by defining a rule that matches the *final* LVD readout *before* compressor failure of the seed truck. It is safe to assume that this is a good positive example, since we always want to flag that vehicle as faulty — in this situation at the latest, and most likely already earlier. Also, given how rich the data we are working with, doing that never covers any other readout. An example of such a rule could be:

$$\text{faulty} \Leftarrow BL = 15.00 \wedge BK = 30.00 \wedge LP = 0.00 \wedge AIZ = 166.53 \wedge ...$$

The next step, then, is to relax this rule, by lowering the specificity of its conditions, in order to *do* cover more readouts. Again, we want to take advantage of the structure of our examples, therefore we select another truck to use for that purpose. We identify the *last* LVD readout of this new truck that is not covered by our rule. We than relax each condition in the rule in such a way that it matches both examples — since all our data is numerical, and we are using basic arithmetic operations, this is a well-defined operation. For example, the previous rule could be relaxed into the following one:

$$\text{faulty} \Leftarrow BL \leq 20.00 \wedge BK \geq 27.00 \wedge LP \leq 0.00 \wedge AIZ \geq 166.53 \wedge ...$$

The next question is how to evaluate the resulting rule. The main goal is to make sure each truck will be flagged as faulty some time before the actual compressor failure, but not too early. We do not have crisp definitions of when is it “too early” and “too late”, however, so we have decided to use fuzzy sets based approach. We have defined the quality of prediction on single truck as shown in Figure 1. In this case we are taking into account *prediction horizon*, i.e. the first time that a warning is issued for a given truck: we are assuming that at this point a new compressor would be installed, and thus the classification on the subsequent readouts is not relevant.

The preferred time for issuing the warning is between 2 and 25 weeks before the fault occurs. Less than that means that the replacement operation may not be possible to schedule, but it is of course still better to issue a warning. Anything longer means

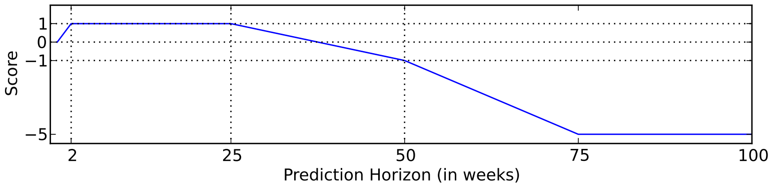


Figure 1. Score function for evaluating classification rules on a single vehicle

that the costs incurred by predictive maintenance scheme start to become significant. In particular, we want to harshly penalise issuing warnings on trucks that never have compressor problems, since those are the most costly mistakes.

The overall rating of a rule is the sum of scores over all previously non-covered trucks. We continue relaxing the rule for as long as the score continues to improve. Once we are done, we remove all trucks correctly classified by this rule from the training set, so that new rules will focus on other examples. We then select a new seed and repeat the whole procedure.

During the rule generation there is one more step we employ. Whenever a rule becomes too generic and covers too many negative examples, we attempt to add a special condition. That condition refers to Vehicle Service Records database, instead of the LVD one, and tests whether one of the repair operations related to the air suspension system has been earlier performed on the given truck.

6. Experiments

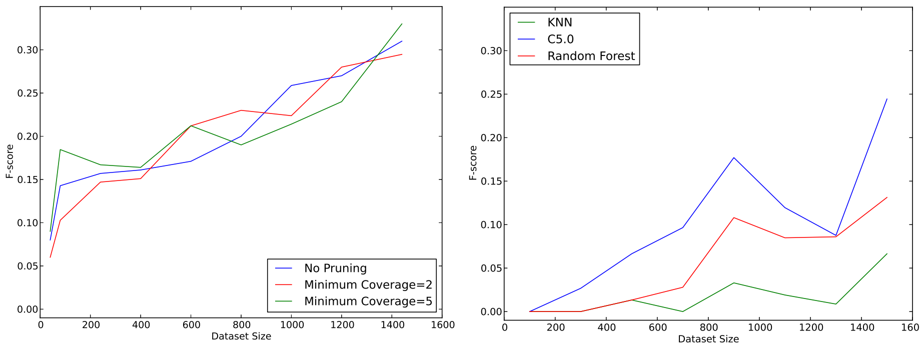
In this section we discuss the experiments we have performed to evaluate the usefulness of the proposed algorithm. We also compare them to results previously described in [12], where we have used a more classical approach. We argue that even those preliminary results demonstrate that RRP algorithm is well-suited to predict compressor failures in trucks. The main reason is that using a per-truck rule evaluation score function rather than binary classification with predefined prediction horizon is better adapted to variance in component wear rate between the vehicles.

We start by showing how the dataset size affects quality of classification. In Figure 2 we compare the F-score as a function of data size for our new algorithm against the results of our previous work. We choose the F-score:

$$F = (1+\beta^2)\frac{\text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}},$$

as this is one of the most popular measures that is suitable for highly imbalanced data sets. We use parameter $\beta = 0.5$, because in this setting precision is more important than recall: compressors that we miss simply maintain *status quo*, while every unnecessary repair costs money. Clearly, classification accuracy is not particularly informative, since the vast majority of examples are non-faulty trucks: a classifier could achieve 95% accuracy by simply answering “no” to all queries.

Figure 2 clearly shows two things. First of all, the dataset we currently have access to is very small, only barely sufficient for the analysis. It is reasonable to assume that,



(a) (b)

Figure 2. Impact of dataset size on $F_{0.5}$-score, comparing (a) the RRP algorithm against (b) results obtained by three popular methods, as described in an earlier work

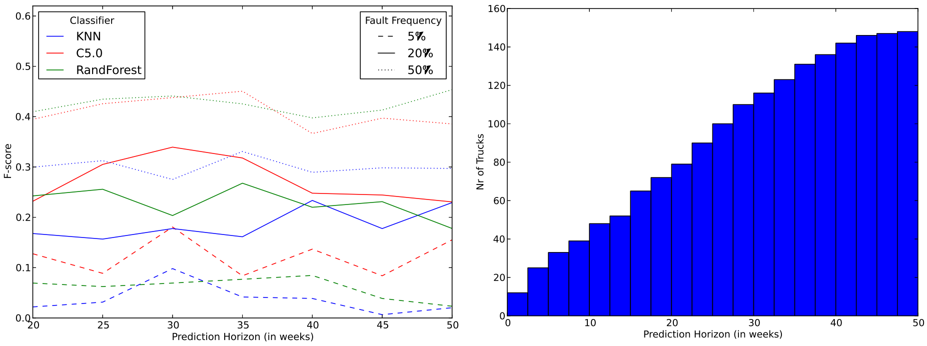
as more data becomes available, the quality of classification will continue to increase. Second, our new algorithm outperforms even the best of classical classifiers, although a direct comparison is not necessarily fair due to different evaluation methodology.

It is also worth mentioning that in Figure 2(a) we present plots for three different levels of a simple rule pre-pruning method. When a new rule is created, if it covers too few examples, we discard it instead of adding it to the solution. At this stage it is difficult to formulate any conclusions concerning usefulness of such pruning scheme.

One reason why we have decided to work on the RRP algorithm was the difficulty in defining a universal prediction horizon for all the trucks. If we treat all data readouts as individual and independent examples, each of them has to be marked as either positive or negative one. All examples closer to the failure than the prediction horizon are considered positive, and all examples further away are considered negative. In Figure 3(a) we show, again from our previous work, how F-score depends on the prediction horizon.

It can be easily seen that there is no “natural” threshold at which prediction horizon should be placed. There is weak or no relation between classification performance and prediction horizon, even when considering datasets with varying fault frequencies. This analysis is also further reinforced by the cumulative histogram presented in Figure 3(b), which shows that when given freedom of choice, the RRP algorithm will issue warnings that are relatively uniformly distributed within the desired time window. This justifies our motivation of moving towards a relaxed prediction horizon, since there is no evidence that a singe, *a priori* specified value would be valid for the whole dataset. On the contrary, it seems that the best prediction horizon is very individual for each truck.

Accuracy and F-score are important measures from the research point of view. The inspiration for our work, however, arises from practical needs of automotive industry, and the major goal from the business perspective is clearly cost reduction. It is very expensive to have components fail during transport missions, because not only does it introduce disruptions in the workshop operations, it also incurs other costs, like towing, collateral damage, and customer dissatisfaction. Therefore, it is much preferred to replace components during scheduled maintenance. The exact degree to which this is the case varies, of course, between components, and also depends on exactly which factors are taken into account — reputation, for example, is notoriously difficult to appraise.

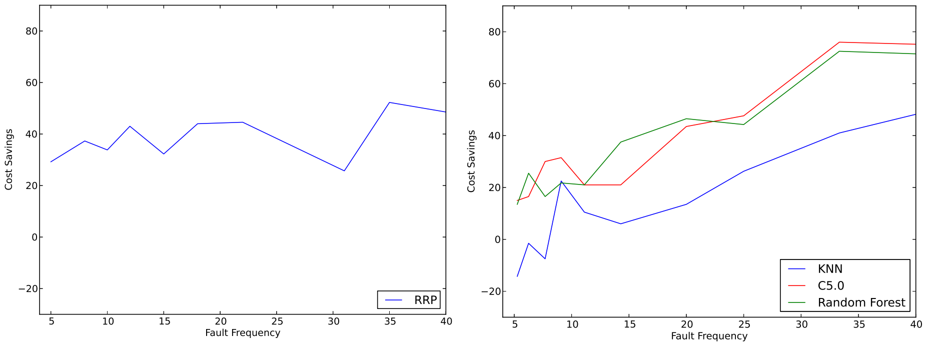


(a) (b)

Figure 3. (a) $F_{0.5}$-scores as a function of predefined prediction horizon and (b) cumulative histogram of fault prediction horizon as learned by RRP

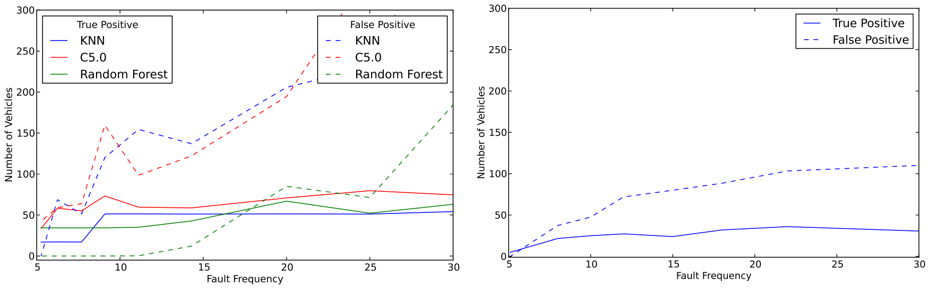
Therefore, in order to be on the safe side, we have decided to use a factor of 2.5 to measure cost savings of predictive maintenance. In other words, we assume that it costs, on average, two and a half as much to repair a truck in which compressor failed on the road, as it would cost to replace this component as a scheduled operation. Figure 4 shows the comparison between RRP and the results obtained in earlier work. As can be seen, our algorithm outperforms all three classical algorithms for highly skewed data sets. This is very important, because the actual fault frequency is in the 5% range (we are not at a liberty to disclose actual value). That said, there is nothing in the design of the algorithm that makes it particularly biased towards skewed data sets, so it would be interesting to analyse why does it not get better as the fault frequency increases. Both KNN and C5.0 clearly outperform RRP before balanced class distribution is reached.

Finally, in Figure 5, we present the relation between True Positives and False Positives, again as a function of fault frequency. This is the plot that actually contains the most information, since those are the two factors that directly affect the economical viability of our solution (presence of False Negatives does not affect the cost in any direct



(a) (b)

Figure 4. Maintenance cost savings, for different fault frequencies.



(a) (b)

Figure 5. True Positives and True Negatives, trained on data with varying fault frequency and tested on data with 5% of faulty trucks

way). It is interesting to notice that the main difference between RRP and all three classical algorithms lies in the fact that the former is significantly more conservative. It issues warnings for the lower percentage of faulty vehicles, but it also makes less mistakes on the healthy trucks. This helps explain the shape of the plot in Figure 4: the cost of false positives is much less important for higher fault frequencies.

It is important to note this was not a conscious design decision, but it is rather a by-product of the way rules are created in RRP. We are not exploring the potential similarity among trucks in any way when choosing the example to be used for relaxing a rule. Also, the syntax we have chosen uniquely determines how each condition will look like after the relaxation process. This means that there is little control over the rule creation, and the whole process is, in practice, rather close to a blind search.

7. Conclusions and Future Work

This work is another step towards a complete, data mining based predictive maintenance solution for automotive industry. The main contribution of this paper is the initial design of the RRP algorithm. It is a novel algorithm with a rich data representation. It uses all observations from a truck as a whole during training, instead of treating each readout individually. It supports our requirement of having a relaxed prediction horizon and does not force any fixed label on a particular readout or observation. With a sharp PH boundary, there is a risk of training on incorrectly labelled data.

The RRP approach, even though the algorithm itself is far from optimised, already offers better classification results and cost avoidance than our previous attempts. Part of this is due to differences in the evaluation methodology, but those differences are justified by the business context. However, we intend to continue working on improving classification quality, for example as we get access to and utilise more data.

In particular, there is a lot of useful information available in the Vehicle Service Records, and we are only using very little of it at the moment. Similarly, the representation of the data in LVD is not perfect. Some form of preprocessing, for instance transforming the data from cumulative and average aggregations to information reflecting the trend between readouts, would likely increase classification performance since correlation between observations would decrease.

Ideas for future work include extending this analysis to other components, especially the ones where "exchange once in a lifetime" assumption does not hold, as well as evaluating known methods of dealing with imbalanced data sets. Additionally, we would like to broaden the data representation in RRP even more, to combine all the relevant data collected from a single vehicle.

RRP should also perform parameter selection automatically, to decrease the dependency on expert knowledge. It is known that data availability will dramatically increase as the new generation of Volvo trucks reach customers. They are equipped with enhanced telematics platform, enabling larger and more frequent LVD readouts. This requires us to develop variable selection algorithms to automatically find the relevant parameters for each vehicle component.

Another natural extension of the work would be to investigate different regression methods to predict *time to repair*. As we mentioned earlier, an important aspect of predictive maintenance in the automotive industry is to be able to capture discrete events that have dramatic effect on the remaining useful life of a component, but combining approaches like RRP with regression seems to be a promising direction.

Acknowledgement

Parts of this work have been supported by Halmstad University, Volvo GIB-T, VINNOVA (the Swedish Governmental Agency for Innovation Systems) and The Knowledge Foundation (KK-stiftelsen).

References

[1] Buddhakulsomsiri, J., Zakarian, A.: Sequential pattern mining algorithm for automotive warranty data. Computers & Industrial Engineering 57(1) (2009)

[2] Byttner, S., Rögnvaldsson, T., Svensson, M.: Consensus self-organized models for fault detection (COSMO). Engineering Applications of Artificial Intelligence 24(5), 833–839 (2011)

[3] Choudhary, A., Harding, J., Tiwari, M.: Data mining in manufacturing: a review based on the kind of knowledge. Journal of Intelligent Manufacturing 20, 501–521 (2009)

[4] D'Silva, S.: Diagnostics based on the statistical correlation of sensors. Tech. Rep. 2008-01-0129, Society of Automotive Engineers (SAE) (2008)

[5] Džeroski, S.: Relational data mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 887–911. Springer US (2010)

[6] Gusikhin, O., Rychtyckyj, N., Filev, D.: Intelligent systems in the automotive industry: applications and trends. Knowledge and Information Systems 12, 147–168 (2007)

[7] Kargupta, H., et al.: VEDAS: A mobile and distributed data stream mining system for real-time vehicle monitoring. In: International SIAM Data Mining Conference (2003)

[8] Khosravi, H., Bina, B.: A survey on statistical relational learning. In: Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence. pp. 256–268 (2010)

[9] Lacaille, J., Come, E.: Visual mining and statistics for turbofan engine fleet. In: IEEE Aerospace Conference (2011)

[10] Michalski, R.S.: Pattern recognition as rule-guided inductive inference. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 349–361 (1980)

[11] Muggleton, S.: Inductive Logic Programming. Morgan Kaufmann (1992)

[12] Prytz, R., Nowaczyk, S., Rögnvaldsson, T., Byttner, S.: Analysis of truck compressor failures based on logged vehicle data. In: International Conference on Data Mining (2013)

[13] Vachkov, G.: Intelligent data analysis for performance evaluation and fault diagnosis in complex systems. In: IEEE International Conference on Fuzzy Systems. pp. 6322–6329 (2006)

[14] Zhang, Y., Gantt, G., et al.: Connected vehicle diagnostics and prognostics, concept, and initial practice. IEEE Transactions on Reliability 58(2) (2009)