Mechanical Systems and Signal Processing 60-61 (2015) 316–325

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Classifying machinery condition using oil samples and binary logistic regression

J. Phillips [a], E. Cripps [b], John W. Lau [b], M.R. Hodkiewicz [a,*]

[a] School of Mechanical and Chemical Engineering, University of Western Australia, Perth 6009, WA, Australia
[b] School of Mathematics and Statistics, University of Western Australia, Perth 6009, WA, Australia

ARTICLE INFO

Article history:
Received 21 June 2013
Received in revised form
14 November 2014
Accepted 27 December 2014
Available online 24 February 2015

Keywords:
Logistic regression
Classification
Oil analysis
Mining trucks
Machine health
Neural networks
Support vector machine
Receiver operating characteristic curve

ABSTRACT

The era of big data has resulted in an explosion of condition monitoring information. The result is an increasing motivation to automate the costly and time consuming human elements involved in the classification of machine health. When working with industry it is important to build an understanding and hence some trust in the classification scheme for those who use the analysis to initiate maintenance tasks. Typically "black box" approaches such as artificial neural networks (ANN) and support vector machines (SVM) can be difficult to provide ease of interpretability. In contrast, this paper argues that logistic regression offers easy interpretability to industry experts, providing insight to the drivers of the human classification process and to the ramifications of potential misclassification. Of course, accuracy is of foremost importance in any automated classification scheme, so we also provide a comparative study based on predictive performance of logistic regression, ANN and SVM. A real world oil analysis data set from engines on mining trucks is presented and using cross-validation we demonstrate that logistic regression out-performs the ANN and SVM approaches in terms of prediction for healthy/not healthy engines.

1. Introduction

This article argues the advantages of using logistic regression (LR) for the binary classification problem of machine health, with an emphasis on a particular industry application that employs an oil analysis based maintenance strategy. Many organisations have embraced oil analysis sampling as part of their condition based maintenance strategy and collect hundreds of samples a month. This process of collection, analysis, and then classification of oil samples by expert analysts is expensive. As one would expect, many of the samples presented to the oil analysts indicate a healthy machine. The samples that require further analysis by the expert are those that indicate a machine is not healthy. Such a classification does not necessarily imply maintenance is required immediately but that extra attention should be directed toward that sample/machine. The ability to automatically classify healthy/not-healthy machines early in the process is attractive so that further analyst effort is not wasted examining the results of samples which show no sign of degradation.

To be suitable for use in industry, a binary classification process should be satisfy the following criteria: (1) the procedure should accurately classify the probability of a machine being healthy/not healthy, (2) the procedure should be easy to use and straightforward to update as new condition data becomes available, (3) it should be clear to the analyst, engineer and

* Corresponding author. Tel.: +61 8 6488 7911.
E-mail address: Melinda.hodkiewicz@uwa.edu.au (M.R. Hodkiewicz).

http://dx.doi.org/10.1016/j.ymssp.2014.12.020
0888-3270/

maintenance planner how changes in explanatory variables affect the probability of the machine being healthy/not healthy and these relationships should make sense in terms of the failure modes experienced by the machine, (4) there should be an ability to assess consequences of misclassification—the cost of catastrophic failure, for example, of these engines can exceed $100,000.

Artificial neural networks (ANN) and support vector machines (SVM) have been thoroughly examined for machine condition classification and are demonstrably capable of addressing points (1), (2) and (4) above. However, point (3) is less accessible. Although labelling such classification schemes as "black boxes" due to their complex structure is perhaps overly strong [1], they definitely present a greater challenge to non-specialists who wish to obtain more insight from the analysis than simply whether a machine is healthy/not healthy and which explanatory variables are the most important. In contrast, LR is well equipped to provide additional insights. The parametric form of LR not only permits the classification probabilities to be estimated for certain levels of explanatory variables, but also how the rates of change of the input variables impact the probability classification. Graphical representations of these relationships are trivial to obtain and, we found, quite informative for the oil analysts with whom we interacted. Also, use of the quantification of misclassification costs via sensitivity, specificity and receiver operating curves provide a mechanism with which to tune the appropriate probability classification level. In light of this, it is surprising that LR has relatively few published examples in machine condition classification. For example, in this journal there have been only two papers [2,3] focussing specifically on logistic regression models, compared to over 20 papers each on ANN and SVM.

Binary LR models have been used for health assessment of an elevator door system [4], light-emitting diodes [5], machinery and cutting tools [2,6], wind turbine bearings [7], helical gear box [8], and simple process systems [9]. Raza et al. [10] used LR to assess the health of a strainer located at the suction side of a pump and compared the results with ANN and support vector machine techniques. Spezzaferro [11] applied LR to aircraft maintenance data and hence determine maintenance inspection interval lengths. Liao et al. [12] compared LR to a proportional hazards model for bearing degradation. LR has been used to estimate the degradation of units (bearings) and more complex facilities and, in combination with a vector machine model, to predict failure probabilities [3,13]. Liu et al. [14] used a combination of a number of methods including LR to assess the performance of a centrifugal compressor turbine. In more recent work Rai [15] compared the classification performance of multinominal LR with decision trees and random forests to predict drill-bit breakage and found that LR had the lowest proportion of classification errors. Also Pandya et al. [16] determined that multinomial LR is more effective than ANN and SVM on classifying bearing faults. With the exception of the work of [10,15,16] none of the other papers compared the performance of their LR models with other binary classification techniques.

Over the last decade ANN models for classification of machine health and fault diagnosis have received significant attention due to their ability to learn from examples, handle incomplete and noisy data, and deal with nonlinear problems [17,18]. There are different types of ANN models used for supervised and unsupervised learning [19,20]. The feed-forward neural network (FFNN) structure is the most commonly used supervised neural network for machine fault diagnostics and the FFNN with back-propagation training algorithm for pattern recognition and classification [21]. Although back-propagation has produced acceptable results for classifying machine condition [22–29], a significant amount of effort is required to determine the architecture and number of nodes. As a result the cascade-correlation neural network (CCNN) was developed [30]. This algorithm does not require initial determination of the network structure and the number of nodes and the efficiency of the network is not compromised by an excessive number of nodes as can occur in back-propagation [31]. Examples of the use of CCNN for machine condition and fault detection are found in [32,33].

SVMs are linear classifiers that make classification decisions based on a linear combination of the features of the data points [34]. While SVMs and LR both calculate a set of weights for variables based on transformation of the feature space, LR models the probability of outcomes explicitly whereas standard SVMs search for the optimal dividing hyper-plane. Perceived advantages of SVM are the ability to manage high dimensional data and that they do not assume a parametric relationship between model predictors and outcomes. One of the disadvantages is the need to select an appropriate kernel function for transforming the covariates, this is non-trivial for a particular classification task as in practice several alternatives need to be considered and compared by means of cross-validation or other methods [35]. Examples of SVM in machine condition classification include [36–41].

In addition to the scarcity of LR models in the condition monitoring literature is the limited number of comparison studies between LR, ANN and SVM. Table 1 shows that most of the focus on comparing these binary classifier performance in machine condition monitoring has been on ANN and SVM [36,37,42,43] with only two papers comparing of LR, ANN and SVM [10,44]. In comparison, outside the machine condition monitoring sector there are many more examples comparing LR with other binary classification models as shown in the review papers in [45,46]. We present a cross-validation study based on our data obtained from oil analysts in industry to classify condition of engines on mining trucks and conclude that in this case LR outperforms an ANN and SVM model in terms of predictive performance. Considering the additional insights provided by LR, and the foremost importance of predictive capability in classification techniques for decision making, we argue LR deserves a larger recognition of its potential in industrial condition monitoring.

The paper proceeds as follows. Section 2 describes the oil condition data for mining trucks with which we demonstrate LR and perform the cross-validation comparison study. These data sets are obtained directly from industry and are typical of those used for condition-based maintenance strategies obtained in the mining industry. Section 3 describes the LR model, Section 4 outlines the CCNN, a form of neural network model, to which we compare the LR model and Section 5 does the same for SVM. Section 6 compares the predictive performances of LR, CCNN and SVM and describes the interpretability of LR