Selection of Input Parameters for Multivariate Classifiers in Proactive Machine Health Monitoring by Clustering Envelope Spectrum Harmonics.

Ann Smith[1, a], Fengshou Gu[1,b] and Andrew Ball[1,c]

[1] School of Computing and Engineering, University of Huddersfield, HD1 3DH, UK

[a]a.smith@hud.ac.uk [b]F.Gu@hud.ac.uk, [c]A.Ball@hud.ac.uk

Keywords: Reciprocating Compressor; Cluster Analysis; Classifiers; Genetic Algorithms; Relevance Vector Machines;

Abstract.

In condition monitoring (CM) signal analysis the inherent problem of key characteristics being masked by noise can be addressed by analysis of the signal envelope. Envelope analysis of vibration signals is effective in extracting useful information for diagnosing different faults. However, the number of envelope features is generally too large to be effectively incorporated in system models. In this paper a novel method of extracting the pertinent information from such signals based on multivariate statistical techniques is developed which substantialy reduces the number of input parameters required for data classification models. This was achieved by clustering possible model variables into a number of homogeneous groups to assertain levels of interdependency. Representatives from each of the groups were selected for their power to discriminate between the categorical classes. The techniques established were applied to a reciprocating compressor rig wherein the target was identifying machine states with respect to operational health through comparison of signal outputs for healthy and faulty systems. The technique allowed near perfect fault classification. In addition methods for identifying seperable classes are investigated through profiling techniques, illustrated using Andrew's Fourier curves.

Introduction

This research investigates the application of multivariate statistical techniques for input parameter selection in data classification modelling. Typically a plethora of potential variables is generated which if utilised cart blanche both obscures salient features and vastly increases run time.

Models of data with a categorical response (e.g. the differing states of health in a gas turbine) are called classifiers. A classifier is built from training data, for which classifications are known. The classifier assigns new test data to one of the categorical levels of the response. Experimentally this is achieved by randomly partitioning data into a training group and one or more test groups.

Reciprocating compressors (RC) are an intrinsic part of many industrial processes whose performance and efficiency rely on early detection of RC component deterioration. There have been many attempts at early diagnosis and classification of RC faults based on vibro-acoustic measurements. These have used a wide variety of modelling techniques for example support vector machines (SVM), radial based functions and relevance vector machines (RVM) [7], all of which focused on improving algorithmic variable manipulation during construction of the classifier rather than pre assessing input variable characteristics.

Prognostics is an emergent capability of current health monitoring which refers to the production of bounded estimates for the remaining useful life of a component or system [2],[3]. An essential feature of an efficient prognostic system is that maximum variability is accounted for by the most sensitive but sparsest model acheivable.

For the analysis discussed in this paper, output signals were collected from the accelerometers on a two-stage, single-acting Broom Wade TS9 RC, which has two cylinders in the form of a "V". Seven output signals were collected including the second stage vibration and motor current signals analysed here. The RC was operated under healthy conditions and with four independently seeded faults (suction valve leakage (SVL), discharge valve leakage (DVL), intercooler leakage (ICL) and loose

drive belt (LB)), each run being repeated 24 times. Thus a total of 120 observations were recorded at each of six pressure loads.

Envelope Spectra Features

Prior research [7] has shown that features extracted from envelope spectra in the frequency domain have superior deterministic properties over their time domain equivalents in monitoring the condition of RCs.. Envelope spectra harmonics exhibit a number of discrete components mainly due to the fundamental frequency of the system (7.3 Hz, the shaft rotation frequency, for the experimental compressor rig employed) and associated harmonics. The fast Fourier transform (FFT) of the envelope signals from vibration sensors were calculated by applying 87835 point with Hanning windows. The magnitude of the FFT being taken as the amplitude or FT spectrum in the spectrum analysis. Envelope spectra show only the amplitude profile of original signals and so provide a clearer insight into the underlying behaviour [12],[13] of the compressor.

In [7] faults were modelled using relevance vector machines (RVM) both with and without genetic algorithm (GA) feature selection. Subsequent classifiers proved highly efficient. Classifiers using multiclass RVM methods were also very successful in correctly classifying cases. However, since Tippings algorithm [8] [9] needs high computional efforts for large numbers of input variables the input feature set was accordingly reduced to include only the envelope features 2 to 15. Whilst dominant features could be identified for inclusion in models the feature selection was more intuitive than quantitative. Clearly features have particular relevance to a given class being repeatedly used in its presence. Moreover features are not uncorrelated so to achieve an optimally sparse model, duplications through inclusion of homogeneous features should be avoided. Mindful, of course, that a little used feature may still provide vital additional information in specific cases.

Analysis of feature characteristics for homogeneity should allow a reduction in the number of input variables (features) through selection of a smaller number of heterogeneous features and elimination of those with contributions to variability already accounted for by others. Thus by applying the underlying model building principles of sparse multivariate regression analysis, reduced feature sets with optimal explanatory powers should be realised.

Cluster Analysis

Cluster analysis (CA) creates groups or clusters of data, features here. Clusters are formed in such a way that objects in the same cluster are very similar and objects in different clusters are distinct. Measures of similarity depend on the application, here Euclidean distance is used. Whilst there are many different CA algororithms there are two main groups i.e. aglomeration techniques whereby all objects start in a group of one and division whereby all objects originate from a single group. Aglomeration was chosen here as the main focus was to identify variable similarity. A hierarchic method being applied to produce a dendrogram for easy visual group identification. Further discussion of CA techniques and applications is given in [10],[11]. Thus CA gives a clear measure of variable connections and dissimilarities whereas RVM and support vector machines (SVM) do not give a clear indication of variable properties. Hierarchical Clustering groups data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy. Investigations into feature similarities through clustering are displayed in the dendrogram Figure 1. A classification matrix, X (120 by 32), was constructed. This comprised 120 observations (24 repeated runs of all 5 classes or machine states) with the first 32 harmonic features being extracted from the envelope spectrum analysis. Dendrograms were based on the pairwise Euclidean distance between each of the observations i.e. a differences vector of length m(m-1)/2 = 120*119/2=7140. Pairwise Euclidean difference between the i^{th} and j^{th} observations, d_{ij}, being given by

$$d_{ij}^{\,2} = (x_i - x_j)(x_i - x_j)' \qquad (1)$$

Thus a square matrix of order m is generated with each entry (i, j) being the Euclidean distance between the observations i and j. From this an agglomerative hierarchical cluster tree was created

using the 'average' linkage method. Average linkage being calculated from the average distance between all pairs of objects in any two clusters: with n_r the number of objects in cluster r, x_{ri} the i^{th} object in cluster r and x_{sj} the j^{th} object in cluster s

$$d_{(r,s)} = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \quad (2)$$

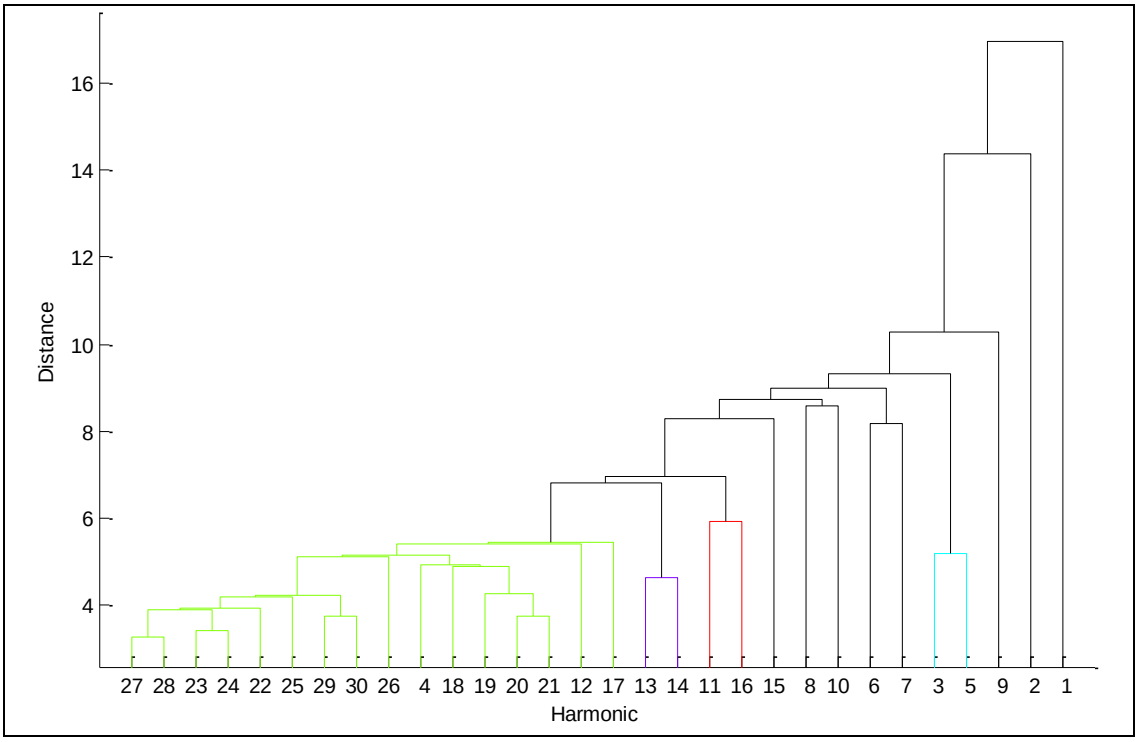

FIGURE 1 SECOND STAGE VIBRATION ENVELOPE FEATURE CLUSTERING BASED ON EUCLIDEAN DISTANCE BETWEEN PAIRWISE OBSERVATIONS.

Clustering of Envelope Features from Second Stage Vibration

The output signals from the second stage vibration and motor current measurements were simultaneously collected from the same compressor rig and so represent precisely the same machine characteristics i.e. healthy or faulty machine states. However, the envelope harmonic features extracted from each spectrum are not identical and the same harmonics are not expected to explain variation for each of the analyses. Neither would the same set of harmonic features be selected as the input parameter set for model classifiers.

A summary of the second stage vibration cluster results is given in Table 1 Features 27, 28 and 23, 24 were found to be most similar whilst the group containing features 3 and 5 is least like all others. There are several other early groupings of features which would be expected to explain similar variation between cases. Other features would also appear to readily form group pairs. For example, features 13 and 14 and features 11 and 16. Interestingly many of these early grouping are with features within the range selected in [7] which would suggest discriminating power may have been lost due to duplication.

	Group Members
Group 1	4 selected as representative of large group of like features.
Group 2	(13,14), (11,16)
Group 3	(3, 5) 6, 7, 8, 10, 15
Independent features	1, 2, 9 (1 omitted as has been shown to have little discriminating power).

TABLE 1 SECOND STAGE VIBRATION ENVELOPE FEATURE CLUSTER GROUPS (GROUP THRESHOLD SET AT 6.0 WITH SUB-SETS SHOWN IN BRACKETS FORMING PRIOR TO THIS DISTANCE)

Extending the threashold to 6.0 highlights the existence of 3 main feature groups plus 3 independent features. The five features joing group 3 at T>14 suggests they are heterogeneous and

potentially provide unique information with respect to separation of cases. Envelope features are grouped by similarity thus forming homogeneous sets of features. A representative feature was selected from each group through inspection of the envelope spectra with the criteria of maximum case separation. Group 3 had a number of late joining features which can be considered as partial outsiders, shown outside the group brackets, and as such are likely to provide additional explanatory power to the model.

It is reasonable to expect the variation within each group might be explained by one representative feature from that group since within group characteristics are homogeneous. Between group differences confirm findings from prior research [7] showing cases to be associated with specific harmonic features.

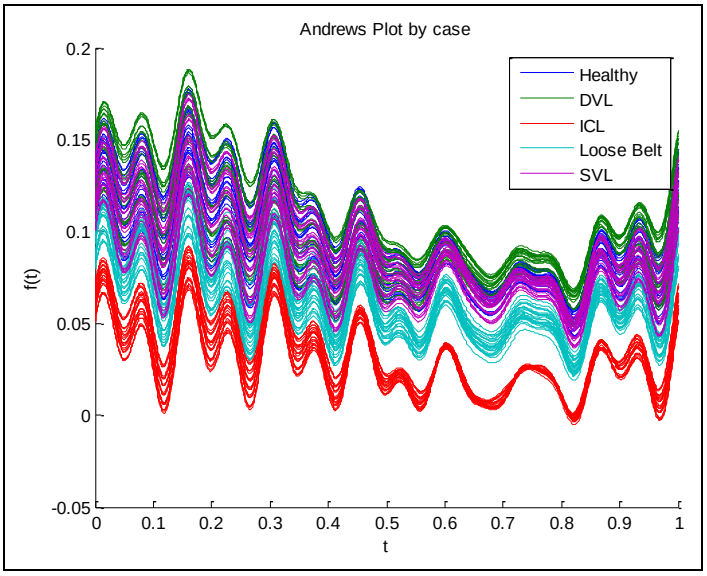

FIGURE 2 ANDREW'S PLOT DISPLAYING CLASS PROFILES FOR THE MOTOR CURRENT OUTPUT SIGNAL.

Andrews Plots

Having identified suitable input parameters for the model building process it is sensible to consider the feasibility of class seperation. One useful profile method is found in Andrew's plots, a Fourier transform of the signal data. This exploratory data analysis technique attempts to identify structure within the data. If there are distinct data groupings, e.g. the clases examined on the compressor rig, then a Fourier profile plot may highlight differences and so assist in distinguishing between groups (classes). The Andrews plots using the motor current signal (Figure 2) show mostly distinct profiles for each of the classes, it would thus be expected models could be constructed with high classification capabilities. The most closely woven classes the DVL and SVL along with the Healthy system which would be more troublesome to differentiate between. On the other hand the ICL profile suggests a far more straightforward classification should be achievable.

Classification Success

Naive Bayes classification using the 11 parameter model suggested by the variable clustering gave near perfect classification across all five categorical classes 3. Model classifiers are further investigated in [14]

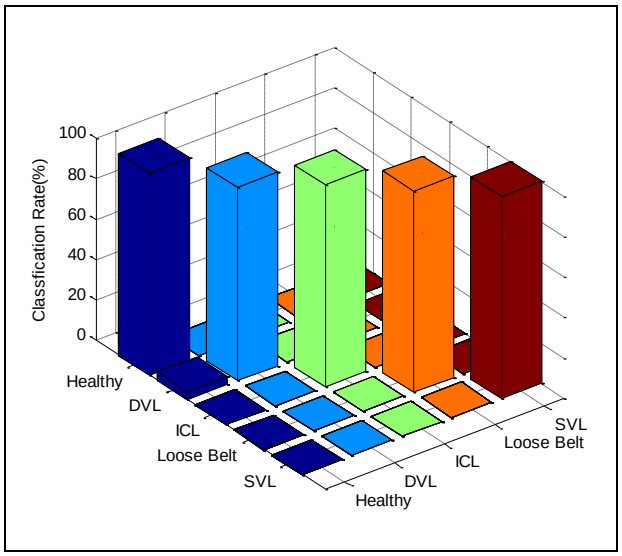

FIGURE 3 SECOND STAGE VIBRATION 11 PARAMETER MODEL WITH 99.2% SUCCESSFUL CLASSIFICATION ACROSS ALL 5 MACHINE STATES.

Conclusions

Investigation of variable interdependencies though data clustering techniques provides a robust method for exploratory variable analysis and so facilitates reduction in the number used in establishing classifiers. A major advantage in condition monitoring in that optimum feature sets are reliably selected prior to incorporation in model algorithms. Reducing run time and ensuring convergence whilst maintaining explanatory power. In turn the algorithm sensitivity with respect to its' ability to detect change is maximised thus increasing opportunities for intervention and health prognosis.

Group profiling e.g. through the use of Andrew's Fourier plots gives an indication of the potential for class seperation thus a measure of the likely classification success of any subsequently developed classifier.

Model efficiency is optimised through clustering potential input parameters with respect to similarities leading to enhanced understanding of variable characteriscs so definining an optimal set of variables for compressor fault classification. However, since clustering gives a clear measure of variable properties this technique is generalisable to different machine monitoring applications.

References

[1] S King, P R Bannister,D A Clifton, and L Tarassenko '*Probabilistic approach to the condition monitoring of aerospace engines'* IMechE Vol. 223 Part G: J. Aerospace Engineering (2009)

[2] M. A Zaidan, A. R. Mills R.F. Harrison '*Bayesian framework for aerospace gas turbine engine prognostics*' IEEE Aerospace Conference Proceedings (2013)

[3] O.W. Laslett, A.R. Mills, M.A. Zaidan, R.F. Harrison '*Fusing an ensemble of diverse prognostic life predictions'*, Aerospace Conference IEEE (2014), p. 1-10

[4] F. Gu, A.D. Ball *'Use of the smoothed pseudo-Wigner-Ville distribution in the interpretation of monitored vibration data Maintenance'* vol. 10 (1995), p. 16-23

[5] B. S. Yang, W. W. Hwang, M. H. Ko, S. J. Lee *'Cavitation detection of butterfly valve using support vector machines.'* Journal of Sound Vibration, vol. 287 (2005), pp 25-43.

[6] B. S. Yang, W. W. Hwang, D. J. Kim, A. Chit Tan *'Condition classification of small reciprocating compressor for refrigerators using artificial neural networks and support vector machines.'* , Mechanical Systems and Signal Processing, vol. 19 (2005), p. 371-390.

[7] M. Ahmed, A. Smith, F. Gu, A.D. Ball '*Fault Diagnosis of Reciprocating Compressors Using Revelance Vector Machines with A Genetic Algorithm Based on Vibration Data*' Proceedings of the 20th International Conference on Automation & Computing **(**2014)

[8] M.E. Tipping '*Sparse Baysian learning and the relevance vector machine.*' The journal of Machine Learning Research, vol. 1(2001), p. 211-244

[9] M.E.Tipping, A. Faul**,** '*Analysis of sparse Baysian learning*' Advances in neural information processing systems, vol. 14 (2002), p. 383-389

[10] Manley, Bryan F. J. *, Multivariate Statistical Methods: A Primer'* 3rd Edition. Chatfield and Colins ISBN 9781584884149. (1986)

[11] B. S. Everitt, G. Dunn *'Applied Multivariate Data Analysis', 2nd Edition.* Arnold, London, ISBN 0 340 74122 8 (2001)

[12] Rao, B. Handbook of Condition Monitoring. Elsevier Science Ltd, Oxford, UK, ISBN 1 85617 2341 (1996)

[13] Isermann**,** R., Fault-Diagnosis Applications. Model-based condition monitoring; Actuators, drives, machinery, plants, sensors, and fault-tolerant systems. Springer. ISBN 978-3-642-12766-3 Springer Berlin Heidelberg New York. (2001)

[14] A. Smith, F. Gu, A.D. Ball *Sparse Statistical Modelling in Prognostic Health Monitoring: Discriminant Analysis and Naive Bayes Classifiers Established Through Input Parameter Selection By Variable Clustering.'* ICAC'15 University of Strathclyde,Glasgow, UK, 11-12 September 2015 (2015 Unpublished),