

High-dimensional visualisation for novelty detection

David A. Clifton, Peter R. Bannister, and Lionel Tarassenko
Department of Engineering Science, University of Oxford
Oxford, OX1 3PJ, U.K.
+44 (0) 1865 273113
davidc@robots.ox.ac.uk

Lei A. Clifton
Nuffield Department of Anaesthetics, University of Oxford
Oxford, OX3 9DU, U.K.

Srini Sundaram
Oxford BioSignals Ltd.
174, Milton Park, Abingdon, Oxfordshire, OX14 4SE

Steve King
Rolls-Royce PLC
Derby, DE24 8BJ, U.K.

Abstract

A key step in the application of novelty detection techniques to high-dimensional data is the exploration of relationships that are typically not obvious when dimensions are examined independently. Visualisation techniques allow such exploration of the structure of the data set by mapping high-dimensional data into lower dimensionality for inspection. This paper discusses the application of the NeuroScale visualisation method for construction of this mapping, which is commonly employed due to its ability to interpolate between training examples into areas of data space not previously encountered. 1) We show that there are disadvantages to using this visualisation method for extrapolation, as is commonly performed when visualising previously-unseen test data which are “abnormal”. 2) We describe a method for ensuring consistent projection of such previously-unseen “abnormal” examples. 3) We show how the proposed technique can also be used to provide a visualisation of high-dimensional decision boundaries, as are typically applied to models of normality in high-dimensional novelty detection cases. An example from a probabilistic model of normality is presented, in which a decision boundary is computed using Extreme Value Statistics and then visualised in two dimensions, showing how the method can be used to communicate the results of novelty detection and allow analysis of high-dimensional “abnormal” data.

1. Introduction

Novelty detection is the construction of a *model of normality* using examples of “normal” system behaviour, and then detecting deviations away from that model, which are classified “abnormal”. This approach is particularly useful when the number of

examples of abnormal behaviour is few, such as would be required to construct a conventional multi-class classifier. In the condition monitoring of high-integrity systems, the reliability of the system means that the majority of data available for constructing models typically correspond to normal system behaviour. Examples of failure are often very rare, and failure modes are often ill-defined due to high system complexity. By adopting a novelty detection approach, this large quantity of normal data, and lack of well-defined fault states, can be exploited.

Visualisation is the process of projecting high-dimensional data (as is typically obtained in condition monitoring of multi-sensor high-integrity systems) into two or three dimensions, suitable for visual inspection. This is an important tool for use during model construction, allowing structure in high-dimensional datasets to be examined and the results of different feature extraction, feature normalisation, and modelling approaches to be compared⁽¹⁾.

Furthermore, visualisation is useful in conveying the results of high-dimensional analysis to a user during monitoring of the system. Visualisation can provide graphical explanation of decisions made by the monitoring system, increasing user confidence and allowing a more intuitive interpretation of system state⁽²⁾.

This paper identifies limitations in a visualisation method often used for novelty detection, whereby inaccurate visualisation of high-dimensional structure may be obtained when “abnormal” data are encountered, and proposes a solution to overcome these limitations.

Noting that the high-dimensional model of normality is associated with a high-dimensional decision boundary, separating “normal” and “abnormal” areas of data space, we show how the proposed method allows visualisation of the high-dimensional decision boundary in two dimensions.

The method is applied to condition monitoring of a modern gas-turbine aerospace engine, where we show that events occurring during engine operation are automatically identified, and correctly presented to the user for graphical interpretation. The proposed method is used to show the high-dimensional decision boundary in two dimensions, providing for the first time an intuitive understanding of novelty detection for a user.

2. Visualisation for Novelty Detection

This section describes Sammon’s mapping and NeuroScale for visualisation of high-dimensional datasets in two or three dimensions.

2.1 Preparation of data

Prior to visualisation, features are extracted from data that best characterise the differences between normal and abnormal system behaviour. Extracted features are typically combined in a (high-dimensional) feature vector for further analysis. While

the process of feature extraction is not within the scope of this paper, we note in passing that the process of visualisation is typically used to determine the suitability of candidate feature vectors for separating normal and abnormal system behaviour⁽¹⁾.

Once a set of feature vectors is constructed, it is typically normalised such that each dimension varies over approximately similar ranges, to ensure that features with large absolute values do not dominate features with smaller absolute values. In previous work, we have used component-wise normalisation, a zero-mean unit-variance transformation, where each feature is standardised with respect to its own mean and variance⁽³⁾.

2.2 Sammon's mapping

Sammon's mapping⁽⁴⁾ is a transformation that attempts to best preserve, in the low-dimensional visualisation space (*latent space*), distances between feature vectors in their original high-dimensional space (*data space*). The Sammon stress metric E is defined using the distance d_{ij} between pairs of feature vectors (x_i, x_j) in data space, and the distance d^*_{ij} between the corresponding pair of projected feature vectors (y_i, y_j) in latent space:

$$E = \frac{1}{\sum_{i < j} d_{ij}} \sum_{i < j} \frac{(d_{ij} - d^*_{ij})^2}{d_{ij}} \quad (1)$$

The distance used is typically Euclidean, and the optimisation is typically performed using gradient descent⁽⁵⁾.

A disadvantage of using Sammon's mapping for projection of data is that the mapping is only defined for those feature vectors upon which it was constructed. This may be suitable for exploration of datasets during model construction, but is unsuitable for visualisation during on-line monitoring due to the time taken to construct a new mapping, as required for the visualisation of previously-unseen data.

(N.b., attempts have been made to increase the speed of constructing Sammon's mapping using sparse-data approaches⁽⁶⁾, and to generalise the mapping to include previously-unseen feature vectors⁽⁷⁾.)

2.3 NeuroScale

NeuroScale⁽⁸⁾ parameterises Sammon's mapping using a single-layer radial basis function (RBF) neural network, in which E is minimised during network training⁽⁵⁾. The number of hidden nodes in the single hidden layer is typically selected to be an order of magnitude greater than the dimensionality of the input feature vectors⁽¹⁾. Each node in the hidden layer corresponds to the centre of a radial basis function (typically a Gaussian kernel) in data space, the initial positions of which are set to be those of feature vectors randomly selected from the training set.

Training the network requires two stages: i) the parameters of the kernel functions are set so that they model the unconditional data density, typically using Expectation-Maximisation⁽⁹⁾ or k -means clustering⁽⁵⁾; ii) the output weights are set by optimisation using linear algebra⁽⁵⁾.

The advantage of using NeuroScale for visualisation is that the mapping from high-dimensional data space to low-dimensional latent space is parameterised by the network weights and RBF centres. This allows previously-unseen feature vectors to be mapped into latent space using a NeuroScale network previously trained on other data. In monitoring systems employing novelty detection techniques, this allows the NeuroScale network to be constructed *a priori*, using the same training data from which models of normality are constructed, with on-line data observed during actual monitoring presented to the network for visualisation by the eventual user of the system.

3. Methods

This section identifies the limitations of using NeuroScale for visualisation of data in novelty detection, and proposes a method for overcoming these limitations. The method is also shown to be useful for obtaining a representation of the high-dimensional decision boundary in the (low-dimensional) visualisation.

3.1 Extrapolation with NeuroScale

Typically in the visualisation of data for novelty detection, a NeuroScale network is trained using a training set mostly comprising “normal” feature vectors, due to the rarity of abnormal examples of system behaviour^(1,2,10,11). The mapping parameterised by the resultant network can project feature vectors that are similar to those included within the training set, due to the ability of the network to *interpolate between* training examples. However, when presented with feature vectors that are significantly different to those included within the training set, the network must *extrapolate* to map regions of data space of which little or no knowledge exists. Relying on the RBF neural network to extrapolate to previously-unseen areas of data space (such as “abnormal” areas) can produce poor results, as the activation functions of the RBFs are only well-defined between training examples. This effect can be particularly evident for previously-unseen feature vectors representing abnormal system behaviour, as will be shown later.

Previous work⁽¹²⁾ has identified this limitation of projection networks, and attempted to avoid the problem by restricting the application of the mapping to those points deemed to be “normal” with respect to some density estimate formed from the normal training examples.

3.2 Avoiding extrapolation

In order to avoid the extrapolation of the neural network to previously-unseen areas of data space, we here propose a method for including feature vectors outside the locus of the available “normal” examples. Typically, such visualisation methods are employed

in conjunction with a modelling technique (such as density estimation^(1,10)) that assumes the normal data are generated from some underlying “normal” data distribution. This density estimation often takes the form of a model containing a mixture of similar kernels, which provides an estimate of the unconditional data density $p(\mathbf{x})$ for feature vector \mathbf{x} . A *novelty threshold*, h , is then set on $p(\mathbf{x})$ such that feature vectors for which $p(\mathbf{x}) < h$ are classified “abnormal” (noting that the novelty threshold defines a decision boundary in data space). For models in which the kernels are isotropic, such as a Parzen window model^(1,10), this threshold occurs at a constant radius from each of the kernels. This paper concentrates on illustrating the proposed method using this case, though it is straightforward to extend it models containing kernels with varying covariance matrices, or models in which each kernel can take a different prior weighting (such as a Gaussian Mixture Model).

3.3 Sampling the decision boundary

Taking an isotropic D -dimensional Gaussian kernel, centred at $\boldsymbol{\mu}$ with width σ and a novelty threshold occurring at some radius r from the centre, we aim to create a set of samples at radius r from the centre as shown in Figure 1.

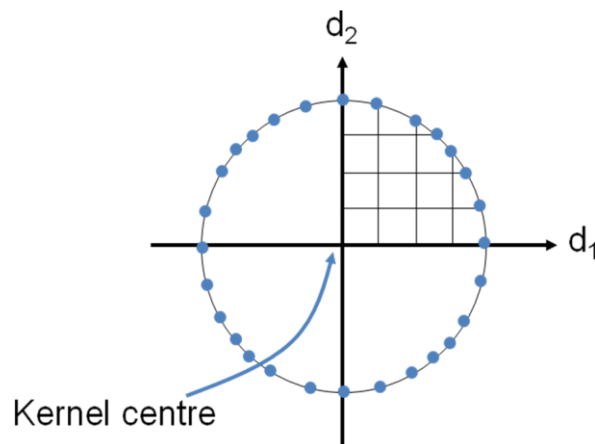


Figure 1 - Sampling a constant radius r around the centre of a bivariate Gaussian kernel. Above, N samples are taken between $(0, r]$ in each dimension $\{d_1, d_2\}$ individually, resulting in $2N$ samples per quadrant, and $4(N-1)$ samples overall

This is achieved using the following algorithm:

```

REPEAT for all dimensions  $d = 1 \dots D$ 
  Create a regular mesh of samples over the remaining  $D-1$  dimensions
  (i.e., a mesh of size  $N^{D-1}$ )
  Project this mesh onto dimension  $d$  using the equation of a (hyper)sphere
  (retaining only those samples which lie on the hypersphere)
END

```

In practice, the symmetry of the isotropic Gaussian kernel can be exploited by first sampling the standard Gaussian ($\boldsymbol{\mu} = 0, \sigma = 1$), sampling each dimension d for $d \geq 0$

(i.e., the upper-right quadrant in the figure), and then reflecting the samples about each axis in turn to provide a sampling around the entire hypersphere. This set of samples from the standard Gaussian can then be mapped onto each kernel i (with centre $\mu_i = 0$ and width σ_i) by scaling all samples by σ_i and translating all samples by μ_i .

As shown in Figure 2, we then retain only those samples that lie on the decision boundary by discarding any point \mathbf{x} for which the closest kernel to \mathbf{x} is not the kernel from which it was generated.

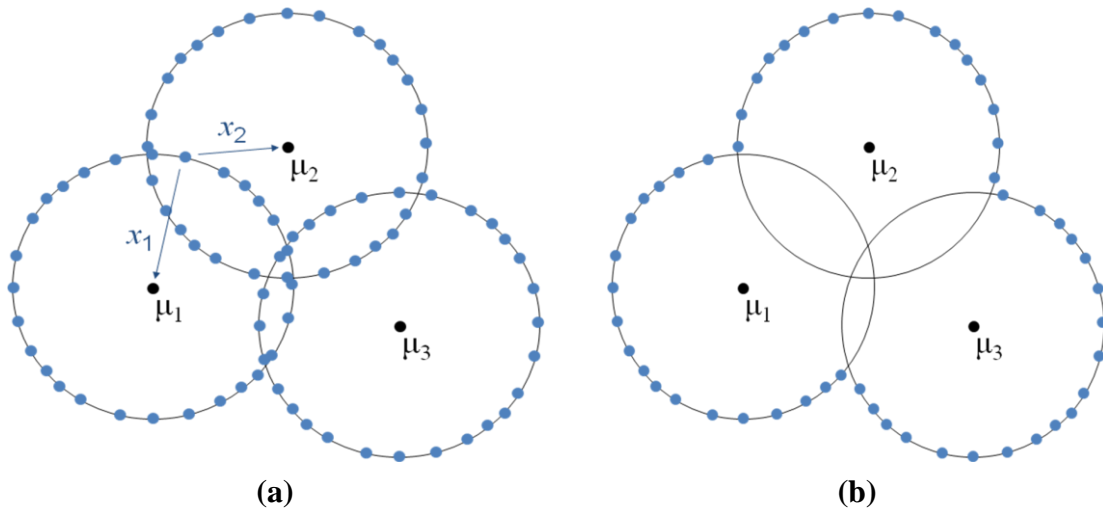


Figure 2 – Removing samples that do not occur on the decision boundary of a kernel-based model. (a) A sample has distances x_1 and x_2 to Gaussian kernels centred at μ_1 and μ_2 , respectively. A sample is retained if the nearest kernel centre is that from which it was generated. Here, the sample would be removed, because it was generated from μ_1 , but $x_2 < x_1$. (b) Results of sample removal, leaving only those on the decision boundary

We note that this construction of a sampling mesh over the D -dimensional hypersphere becomes increasingly difficult for higher D , where the regularity of the sampling grid would need to be exchanged for random sampling of the decision boundary. This could be achieved by sampling the D -dimensional Gaussian distribution, and normalising each sample by σ , such that all samples are mapped to the hypersphere radius r . In practice, we have found the proposed method adequate up to $D = 10$ dimensions.

An example of the proposed approach is shown in Figure 3, in which a model is formed from data with $D = 2$ dimensions, using 500 Gaussian kernels, each centred on one of 500 normal training examples. Here, we have set the novelty threshold to be a small radius r from the centre of each kernel, resulting in a decision boundary that overfits the data.

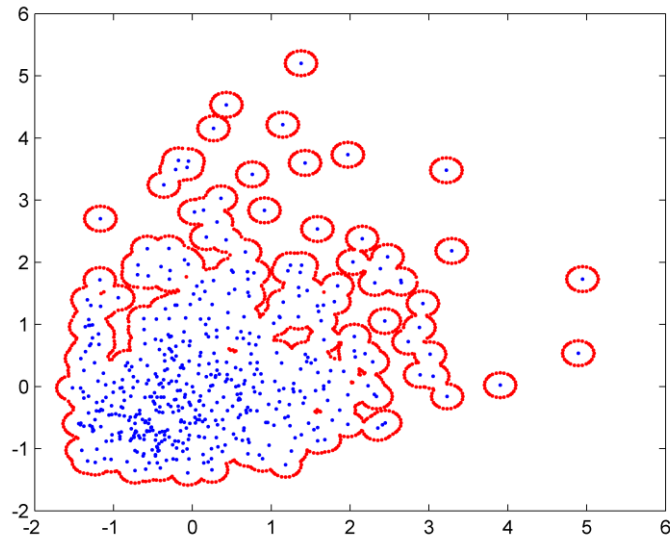


Figure 3 – Example 2-D model formed from 500 Gaussian kernels (shown in blue), with samples from the decision boundary (shown in red). The boundary is piecewise circular

An example for data with $D = 3$ dimensions is shown in Figure 4, again using 500 Gaussian kernels, each centred on one of 500 normal training examples. Here, we have set the novelty threshold to be a larger radius r from the centre of each kernel, resulting in a decision boundary that better describes the distribution of normal data. Methods for setting the optimal width σ in a Parzen windows model for novelty detection are described in (10).

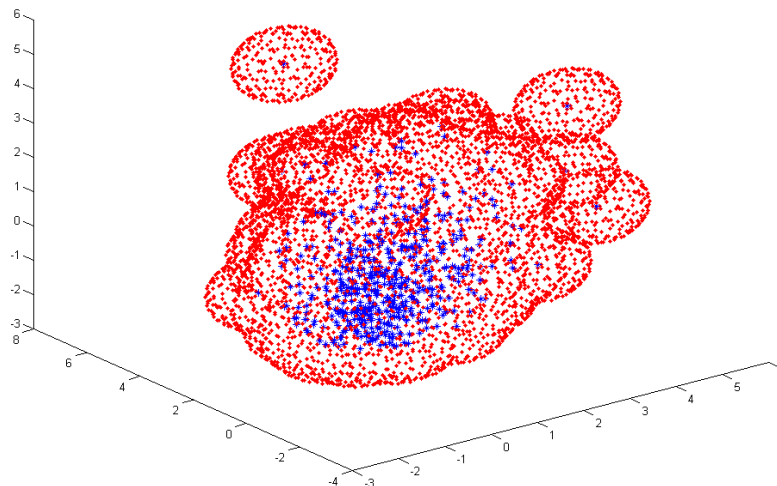


Figure 4 – Example 3-D model formed from 500 Gaussian kernels (shown in blue), with samples from the decision boundary (shown in red). The boundary is piecewise spherical

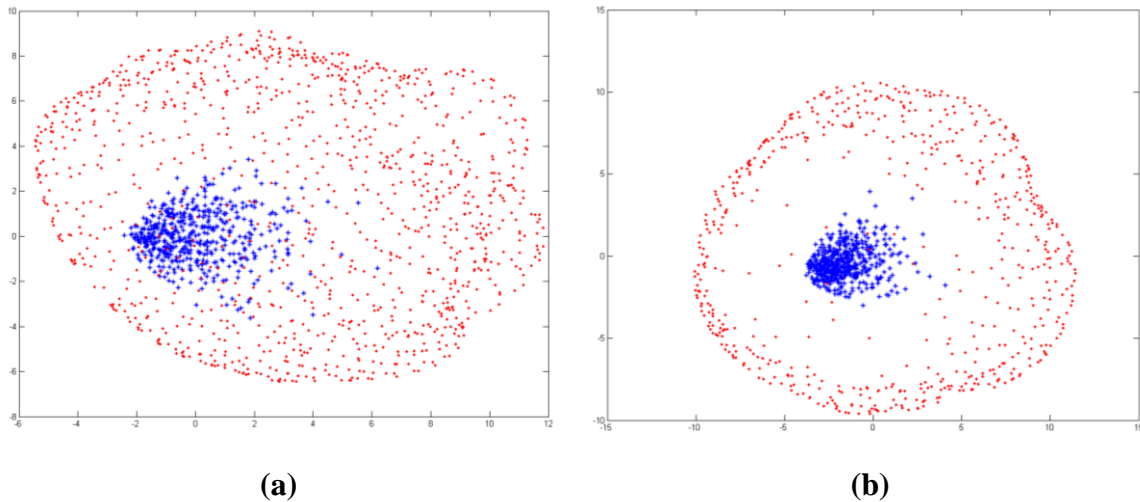


Figure 5 – Example 8-D model projected into 2-D for visualisation, composed of 500 Gaussian kernels (shown in blue), with projections of samples from the 8-D decision boundary (shown in red). (a) Conventional training of the projection using only the model kernel centres. (b) Training the projection using model kernel centres and samples from the decision boundary.

When dealing with data for which $D > 3$, visualisation methods are employed to represent the data in two dimensions. Figure 5(a) shows the result of training a NeuroScale network using only the 500 normal feature vectors. It may be seen that the feature vectors sampled from the 8-D decision boundary appear to be projected throughout the locus of normal data, which is undesirable. Feature vectors from the decision boundary are far from normal data in 8-dimensional space, and should be correspondingly separated from the normal data in two dimensions. This illustrates the limitations of relying on the NeuroScale network to extrapolate its mapping into areas of data space where no training examples are available.

Figure 5(b) shows the result of including feature vectors sampled from the 8-D decision boundary within the training set. The majority of these feature vectors (shown in red) are now correctly projected as being separated from the normal feature vectors (shown in blue). We have avoided the need to rely on extrapolation into “abnormal” areas of data space by including the sampled feature vectors. Thus, we have constructed a visualisation network that can cope with data lying significantly far from the “normal” training examples, and which can be used for correct communication of system state to users during monitoring.

4. Data and Results

This section presents the results of applying the proposed method to actual engine data, and illustrates how the method can be used to display previously-unseen “abnormal” feature vectors in an accurate manner, while also displaying a low-dimensional representation of the high-dimensional decision boundary.

4.1 Gas-turbine engine data

The data used in the investigation described by this paper were obtained from a modern aerospace gas-turbine engine. The dataset consists of vibration amplitudes observed from case-mounted sensors for a single engine over 137 flights. From broadband vibration spectra computed every 0.2s, the amplitude of the various modes of vibration corresponding to the rotation of the engine’s shafts were extracted⁽¹⁰⁾ and used for subsequent analysis.

4.2 Constructing a model of normality

Vibration amplitudes for eight modes of vibration were used to form a $D = 8$ -dimensional feature vector \mathbf{x} every 0.2s.

Vibration amplitudes from “normal” flights [35–79] were reduced using k -means clustering to a set of $k = 500$ prototype cluster centres. A model of normality was constructed using Parzen windows by placing a Gaussian kernel at each of the 500 cluster centres, with global width parameter σ set to be the Euclidean distance between each cluster centre and its 10 nearest neighbours, averaged over all cluster centres.

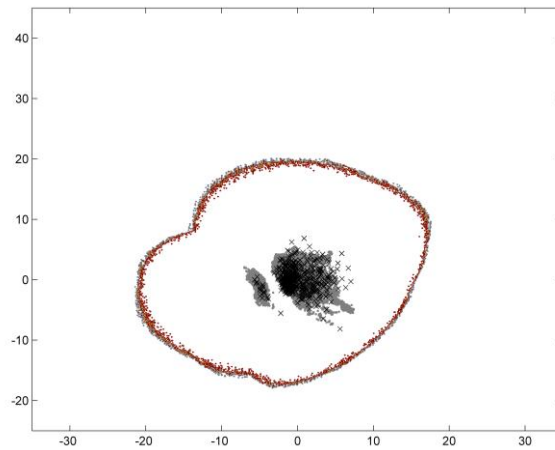
With the probability $p(\mathbf{x})$ now available from the Parzen window model, a novelty threshold h was set on $p(\mathbf{x})$ using a method employing Extreme Value Statistics, of which discussion is deferred to (10). As described previously, this novelty threshold corresponds to a contour on $p(\mathbf{x})$ occurring at some radius r from the Gaussian kernels, which we sample with $N = 7$, retaining only those samples on the decision boundary.

The resultant feature vectors that lie on the decision boundary are combined with the normal training vectors and used to train a NeuroScale network to project from the 8-dimensional data space into two dimensions for visualisation.

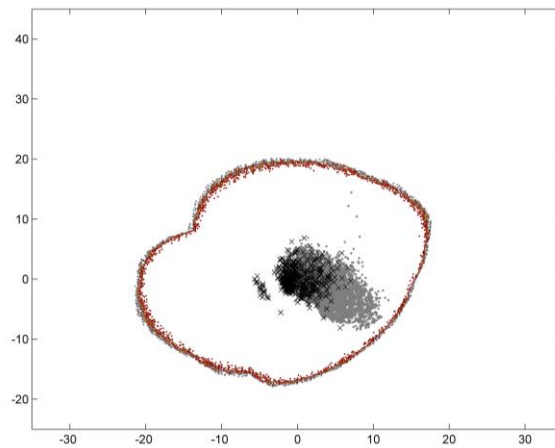
4.3 Visualisation of flight data

Examples of visualised feature vectors from three example tests are shown in Figure 6. In each case, the projected model kernel centres are shown (by black crosses) with the projected samples from the decision boundary (in red), and the feature vectors are shown for each flight (in grey).

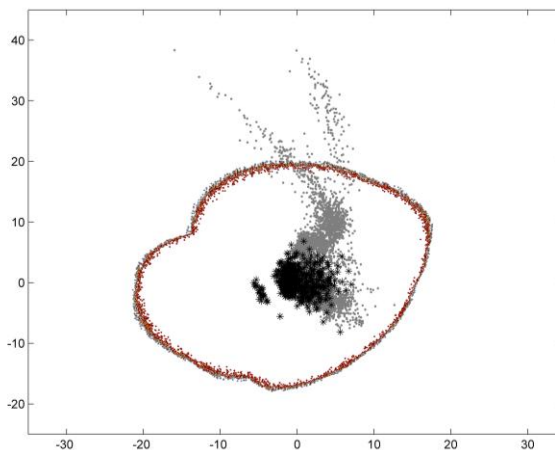
Figure 6(a) shows 2-dimensional projections of 8-dimensional feature vectors from a “normal” flight. It may be seen that the projected feature vectors closely overlay the projected centres from the model of normality. Figure 6(b) shows projections of feature vectors from flight 131, when the condition of the engine has changed due to an internal engine event. The projected feature vectors show significant deviation from the projected model centres, and approach the representation of the decision boundary. Figure 6(c) shows projections of feature vectors from the final flight in the series, in which a major engine event has occurred. The projected feature vectors significantly breach the representation of the decision boundary, providing an accurate image of the 8-dimensional system state. (N.b., actual determination of “abnormality” occurs in high-dimensional space.)



(a)



(b)



(c)

Figure 6 – 2-D projection of 8-D model kernel centres (shown as black asterisks), with test data from a single flight (shown in grey), and samples from the decision boundary (shown in red). (a) Test data from “normal” flight 75. (b) Test data from “abnormal” flight 131. (c) Test data from flight 137, during which an event was observed

5. Conclusions

We have shown that the conventional use of NeuroScale neural networks for projecting high-dimensional data into two or three dimensions can provide unrepresentative visualisations when dealing with data significantly different to the “normal” data used to construct models of normality. We have proposed a method to avoid this limitation, such that accurate visualisation of high-dimensional data (particularly “abnormal” data) is possible. This method also allows a visualisation of the decision boundary from high-dimensional space.

The results from this paper allow visualisation mappings to be constructed *a priori*, such that previously unseen data may be accurately visualised by a user during monitoring, and visually compared to a representation of the high-dimensional decision boundary. The provision of an accurate visualisation of high-dimensional system state is of key importance in graphically communicating the results of multivariate novelty detection to users, and in increasing user confidence in the monitoring system.

We have identified areas for future work, including the extension of the system to more general mixture models (such as full-covariance Gaussian Mixture Models), and in the use of stochastic sampling to visualise datasets of dimensionality $D > 10$.

Acknowledgements

The authors acknowledge the support of Oxford BioSignals Ltd. and the HECToR research programme, funded by the U.K. Department of Trade and Industry. DAC wishes to thank Iain G.D. Strachan, Dennis King, Paul Anuzis, and Robert Slater for valuable discussions.

References

1. D.A. Clifton, L.A. Clifton, P.R. Bannister, and L. Tarassenko, ‘Automated Novelty Detection in Industrial Systems’, in Y. Liu et al. (eds), *Advances in Computation Intelligence in Industrial Systems*, Springer-Verlag, Berlin, 2008.
2. D.A. Clifton, B. Haskins, P.R. Bannister, and L. Tarassenko, ‘Specific and Generic Models for Jet Engine Novelty Detection’, *Proceedings of the 4th IET International Conference on Condition Monitoring*, Harrogate, UK, pp. 478-487, 2007.
3. D.A. Clifton, P.R. Bannister, and L. Tarassenko, ‘Application of an Intuitive Novelty Metric for Jet Engine Condition Monitoring’, in M. Ali et al. (eds.), *Advances in Applied Artificial Intelligence, Lecture Notes in Artificial Intelligence*, Vol. 4031, pp. 1149-1158, Springer-Verlag, Berlin, 2006.
4. J.W. Sammon, ‘A Non-linear Mapping for Data Structure Analysis’, *IEEE Transactions on Computers* 18(5), pp. 401–409, 1969.
5. I. Nabney, ‘*Netlab: Algorithms for Pattern Recognition*’, Springer-Verlag, Berlin, 2002.

6. M. Martin-Merino and A. Munoz, 'A New Sammon Algorithm for Sparse Data Visualization', Proceedings of the 17th IEEE International Conference on Pattern Recognition, pp. 477-481, 2004.
7. E. Pekalska, D. De Ridder, R.P.W. Duin, and M.A. Kraaijveld, 'A New Method of Generalizing Sammon Mapping with Application to Algorithm Speed-Up', Proceedings of the 5th Conference of the Advanced School for Computing and Imaging, Heijen, The Netherlands, pp. 221-228, 1999.
8. D. Lowe and M.E. Tipping, 'Feed-Forward Neural Networks and Topographic Mappings for Exploratory Data Analysis', Neural Computing and Applications, Vol. 4, pp 83-95, 1996.
9. C.M. Bishop, 'Pattern Recognition and Machine Learning', Springer-Verlag, Berlin, 2006.
10. L. Tarassenko, D.A. Clifton, P.R. Bannister, S. King, and D. King: 'Novelty Detection', in K. Worden, et al. (eds): Encyclopaedia of Structural Health Monitoring. John Wiley and Sons, New York, 2008.
11. D.A. Clifton, P.R. Bannister, and L. Tarassenko: 'Novelty Detection in Large-Vehicle Turbochargers', in H.G. Okuno and M. Ali (eds), New Trends in Applied Artificial Intelligence, Lecture Notes in Computer Science, Vol. 4750, Springer-Verlag, Berlin, pp. 591-600, 2007.
12. C.M. Bishop, 'Novelty Detection and Neural Network Validation', IEE Proceedings on Vision, Image, and Signal Processing, 1994.