Novelty Detection in Learning Systems

Stephen Marsland †
† Division of Imaging Science and Biomedical Engineering,
Stopford Building, The University of Manchester,
Oxford Road, Manchester M13 9PL, UK

Abstract

Novelty detection is concerned with recognising inputs that differ in some way from those that are usually seen. It is a useful technique in cases where an important class of data is under-represented in the training set. This means that the performance of the network will be poor for those classes. In some circumstances, such as medical data and fault detection, it is often precisely the class that is under-represented in the data, the disease or potential fault, that the network should detect. In novelty detection systems the network is trained only on the negative examples where that class is not present, and then detects inputs that do not fits into the model that it has acquired, that it, members of the novel class.

This paper reviews the literature on novelty detection in neural networks and other machine learning techniques, as well as providing brief overviews of the related topics of statistical outlier detection and novelty detection in biological organisms.

1 Introduction

Novelty detection, recognising that an input differs in some respect from previous inputs, can be a useful ability for learning systems, both natural and artificial. For animals, the unexpected perception could be a potential predator or a possible victim. By detecting novel stimuli, the animal's attention is directed first to the most potentially dangerous features of its current environment. In this way, novelty detection reduces the large amount of extraneous information that the animal is receiving, so that it can focus on unusual stimuli.

This application of novelty detection could also be useful for a learning system, where the system only learns about inputs that it has not seen before, thus saving resources. Another area where novelty detection is particularly useful is where an important class is under-represented in the data, so that a classifier cannot be trained to reliably recognise that class. Typical examples of this problem include medical diagnosis, where there may be hundreds of test results, but relatively few show the symptoms of a particular disease, and machine fault recognition, where there may be many hours of operation between failures.

These tasks can be thought of as inspection tasks, where correctly recognising every fault is more important than the occasional false alarm. For inspection tasks, novelty detection has another benefit. Even if a classifier has learnt to reliably detect examples of the important class, a variant may occur, or two diseases could display symptoms simultaneously. These will appear different to the trained examples, and could therefore be missed. However, if the classifier has not see any examples of this class, any similar inputs will not be recognised and so will be detected by a novelty filter.

In general, most novelty filters work by learning a representation of a training set that only contains examples of the 'normal' or 'health' data and then attempting to decide whether or not a particular input

[0]This work was supported by a UK EPSRC studentship. Updates, corrections, and comments should be sent to Stephen Marsland at stephen.marsland@man.ac.uk.

differs markedly from the elements of the training set. A variety of methods of doing this have been proposed, from calculating the projection of each new input into the space orthogonal to the principal components of the training set, to examining which nodes in a Self-Organising Map fire for each input.

Most of the novelty filters that are described in this paper are based on batch training methods. A training set that is known to contain no examples of the important class is created and used to train the novelty filter. The filter then evaluates each new input for novelty with regard to the acquired model. However, this is limiting for datasets that are not known in advance, or that change over time. A small number of researchers have investigated solutions to these problems, and where novelty filters are capable of such on-line operation, this is indicated.

Novelty detection is related to the problem of statistical outlier detection, a brief description of which is given in section 2. Then, methods of novelty detection in supervised neural networks are described, beginning with Kohonen and Oja's orthogonalising novelty filter in section 3 and novelty detection with supervised neural networks in section 4. Section 5 describes the gated dipole, a biologically inspired construct that can perform novelty detection. Following this, section 6 describes a number of novelty detection techniques based on self-organising networks, both supervised and unsupervised, while section 7 lists a number of other methods that have been proposed. Finally, sections 8 and 9 provide an overview of some relevant topics in the biology literature.

2 Outlier Detection

2.1 Introduction

The problem of statistical outlier detection is closely related to that of novelty detection. No precise definition of an outlier seems to have been produced, but most authors agree that outliers are observations that are inconsistent with, or lie a long way from, the remainder of the data. Outlier detection aims to handle these rogue observations in a set of data, since they can have a large effect on analysis of the data (datapoints that have a large effect are known as influential observations). The principal difficulty is that it is not possible to find an outlier in multivariate data by examining the variables one at a time.

How important outlier detection is to statistical methods can be seen in figure 1 – an outlying datapoint can completely change the least-squares regression line of the data. Generally, statistical methods are concerned with ignoring unrepresentative data, rather than explicitly recognising those points. Techniques that avoid many of the problems of outliers are known as *robust statistics* (Huber, 1981). The way that robust statistics can be used for outlier detection is described in section 2.3. There are also sets of tests for deciding whether predictions from particular distributions have been affected by outliers, see for example Barnett and Lewis (1994). The appearance of some outliers in two dimensions are shown in figure 2. The next five subsections describe statistical techniques that are used to detect and deal with outlying datapoints. The related problem of density estimation is discussed in section 4.2.

2.2 Outlier Diagnostics

The residual of a point is $r_i = y_i - \hat{y}_i$, that is, the difference between the actual point (y_i) and the prediction of a point ($\hat{y}_i$). The linear model of statistical regression for data $\mathbf{X}$ is:

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e}, \tag{1}$$

where θ is the vector of (unknown) parameters and $\mathbf{e}$ is the vector of errors. The hat matrix, $\mathbf{H}$ (so called because $\mathbf{Hy} = \hat{\mathbf{y}}$) is defined as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{2}$$

Then

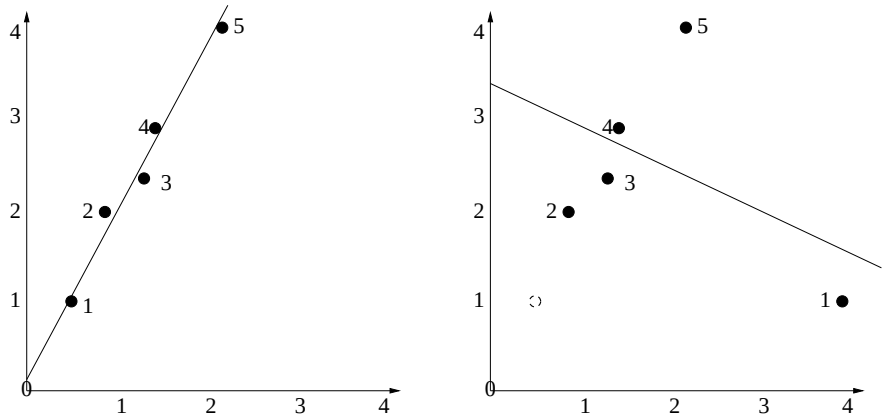

Figure 1: A demonstration of why outlier detection is important in statistics. The five points comprise the data and the line is the least-squares regression line. In the graph on the right, point 1 has been misread. It can be seen that this completely changes the least-squares regression line.

$$\text{cov}(\hat{\mathbf{y}}) = \sigma^2 \mathbf{H}, \tag{3}$$

$$\text{cov}(\hat{\mathbf{r}}) = \sigma^2 (\mathbf{I} - \mathbf{H}), \tag{4}$$

where $\mathbf{r}$ is the vector of residuals and σ^2 the variance. So, each element h_{ij} of $\mathbf{H}$ can be interpreted as the effect exerted by the jth observation on $\hat{y}_i$, and $h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$, the effect that an observation has on its own prediction. The average of this is p/n, where $p = \sum_{i=1}^{n} h_i$, and in general points are considered to be outliers if $h_{ii} > 2p/n$ (Rousseeuw and Leroy, 1987).

It is interesting to note that $\mathbf{H}$ is the pseudoinverse if $(\mathbf{X}^T\mathbf{X})^{-1}$ exists. Therefore, the hat matrix method is related to the approach of Kohonen and Oja (see section 3), which can be considered as an implementation of the hat matrix.

The values along the diagonal of the hat matrix can be used to scale the residuals. Three methods are shown below:

standardised $\frac{r_i}{s}$, where $s^2 = \frac{1}{n-p} \sum_{j=1}^{n} r_j^2$

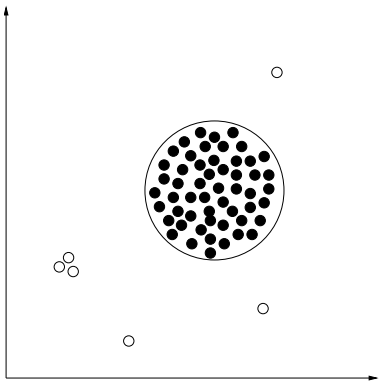

Figure 2: The principle of outlier detection. Empty circles are outliers to the dataset comprised of the black circles. The circle surrounding the datapoints demonstrates a potential threshold, beyond which points are outliers.

studentised $\frac{r_i}{s\sqrt{1-h_{ii}}}$

jackknifed $\frac{r_i}{s_i\sqrt{1-h_{ii}}}$, ($s_i = s$ without the ith case).

Another method that is used is the Mahalanobis distance of each point:

$$D^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{5}$$

where Σ is the covariance matrix and $\boldsymbol{\mu}$ the mean. The Mahalanobis distance is a useful measure of the similarity between two sets of values.

2.3 Robust Statistics

The area of robust statistics is concerned with providing statistical techniques that can operate reliably in the presence of outliers and when the data to be analysed was not necessarily generated by a process that is in the class of theoretical models that underlie the technique. This approach can be used for novelty or outlier detection by highlighting those datapoints that the robust measures ignore or bias against.

The standard texts are Huber (1981), Hoaglin et al. (1983) and Rousseeuw and Leroy (1987) which describe most of the typical approaches, such as those based on maximum likelihood techniques (M-estimates), linear combinations of order statistics (L-estimates) and those based on rank tests (R-estimates). Also of interest is work on robust estimation of the covariance and correlation matrices, see for example Denby and Martin (1979).

2.4 Recognising that the Generating Distribution has Changed

One question that outlier detection aims to answer can be phrased:

Given n independent random variables from a common, but unknown, distribution μ, does a new input X belong to the support of μ?

The support, or kernel, of a set is a binary valued function that is positive in those areas of the input space where there is data, and negative elsewhere. The standard approach to the problem of outlier detection (Hájek and Šidák, 1967) is to take further independent measurements of the new distribution, which is assumed to have a common probability measure ν, and to test if $\mu' \neq \nu$, where μ' is the probability measure of μ, i.e., to see if the support of $\nu \in S$, where S is the support of μ. The problem then is how to estimate the support S from the independent samples $X_1, \ldots X_n$.

The obvious approach (Devroye and Wise, 1980) is to estimate S_n as

$$S_n = \bigcup_{i=1}^{n} A(X_i, \rho_n), \tag{6}$$

where $A(x, a)$ is the closed sphere centred on x with radius a and ρ_n is a number depending only upon n.

Then the probability of making an error on datapoint X, given the data so far, is

$$\begin{aligned} L_n &= P\left(X \in S | X_1, \ldots X_n\right) \\ &= \nu(S). \end{aligned} \tag{7}$$

The detection procedure is said to be consistent if $L_n \rightarrow 0$ in probability, and strongly consistent if $L_n \rightarrow 0$ with probability one.

2.5 Extreme Value Theory

In Roberts (1998), extreme value theory (EVT) (Gumbel, 1958) is used to approach the problem of detecting outliers in data. The approach investigates the distributions of data that have abnormally high or low values in the tails of the distribution that generates the data.

Let $\mathcal{Z}_m = \{z_1, z_2, \ldots, z_M\}$ be a set of m independent and identically distributed random variables $z_i \in \mathbb{R}$ drawn from some arbitrary distribution $\mathcal{D}$, and let $x_m = \max(\mathcal{Z}_m)$. Then, when observing other samples, the probability of observing an extremum $x \geq x_m$ may be given by the cumulative distribution function

$$p(x_m|\mu_m, \sigma_m, \gamma) = \exp\left\{-\left[1 + \frac{\gamma(x_m - \mu_m)}{\sigma_m}\right]^{1/\gamma}\right\}, \tag{8}$$

where $\gamma \in \mathbb{R}$ is the shape parameter. In the limit as $\gamma \to 0$, this leads to the Gumbel distribution

$$P(x_m \leq x|\mu_m, \sigma_m) = \exp\left\{-\exp(-y_m)\right\}, \tag{9}$$

where μ_m and σ_m depend on the number of observations m, and y_m is the reduced variate

$$y_m = \frac{(x_m - \mu_m)}{\sigma_m}. \tag{10}$$

2.6 Principal Components Analysis

Principal Components Analysis (PCA) is a standard statistical technique for extracting structure from a dataset by performing an orthogonal basis transformation to the coordinate system in which the data is described. This can reduce the number of features needed for effective data representation.

PCA can be used for detecting outliers that are in some sense orthogonal to the general distribution of the data (Jolliffe, 1986). By looking at the first few principal components, any datapoints that inflate the variances and covariances to a large extent can be found. However, by looking at the last few principal components, features that are not apparent with respect to the original variables (i.e., outliers) can be seen. There are a number of test statistics that have been described to find these points. Two examples are a measure of the sum of squares of the values of the last few principal components and a version that is weighted by the variance in each principal component. Further details can be found in Jolliffe (1986).

3 Kohonen and Oja's Novelty Filter

3.1 The Novelty Filter

The first known adaptive novelty filter is that of Kohonen and Oja (1976), who proposed an orthogonalising filter that extracts the parts of an input vector that are 'new', with respect to previously learnt patterns. This is the desired functionality of a novelty filter. Another description of the filter is given in (Kohonen, 1993). The Kohonen and Oja novelty filter is a pattern matching algorithm with the following properties:

- patterns that are seen in training are stored
- inputs are compared to each of these stored patterns
- the best-matching stored pattern is selected
- the difference between the best-matching replica and the input is displayed

Mathematically, the effects of the novelty filter can be described as follows. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ be distinct Euclidean vectors spanning a linear subspace $L \subset \mathbb{R}^n$. Then any vector $\mathbf{x} \in L$ can be uniquely written as

$$\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}}, \tag{11}$$

where $\hat{\mathbf{x}} \in L$ is the orthogonal projection of $\mathbf{x}$ on L, and $\tilde{\mathbf{x}} \in L^{\perp}$ is the projection of $\mathbf{x}$ on $L^{\perp}$, the complement space that is orthogonal to L.

It can be shown that the decomposition in equation 11 has the property that $\|\tilde{\mathbf{x}}\|$ is the distance of $\mathbf{x}$ from L, i.e.,

$$\text{norm}\|\tilde{\mathbf{x}}\| = \min_{\|\hat{\mathbf{x}}\|} \|\tilde{\mathbf{x}}\|. \tag{12}$$

Let $A \in \mathbb{R}^{n \times m}$ be a matrix with $\mathbf{x}_i$ as the ith column. Then

$$\hat{\mathbf{x}} = AA^{+}\mathbf{x}, \tag{13}$$

where A^{+} is the pseudoinverse (Penrose, 1955) of A, and so

$$\tilde{\mathbf{x}} = (I - AA^{+})\mathbf{x}, \tag{14}$$

with I being the $2n \times 2n$ identity matrix.

The Gram-Schmidt process can be used to compute the orthogonal projections of vectors. A new vector basis is defined by the following recursion for the subspace L spanned by training vectors $\{\mathbf{x}_i\}, i = 1, \ldots, m$:

$$\tilde{\mathbf{x}}_1 = \mathbf{x}_1, \tag{15}$$

$$\tilde{\mathbf{x}}_k = \mathbf{x}_k - \sum_{i=1}^{k-1} \frac{\mathbf{x}_k \tilde{\mathbf{x}}_i^T}{\|\tilde{\mathbf{x}}_i\|^2} \tilde{\mathbf{x}}_i, \tag{16}$$

where the sum is over the $\|\tilde{\mathbf{x}}_i\| \neq 0$. Then

$$\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_{m+1} \tag{17}$$

and

$$\hat{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}}_{m+1}, \tag{18}$$

so that $\tilde{\mathbf{x}}$ is the residual $\notin L$, i.e., the part of $\mathbf{x}$ that is independent of the vectors $\{\mathbf{x}_i\}$. This is the amount of $\mathbf{x}$ that is 'maximally new', the 'novelty' in $\mathbf{x}$.

So a neural network that has equation 14 as the transfer function will extract the novelty in the input. Kohonen and Oja (1976) propose a network with neurons that implement

$$\tilde{x}_i = x_i + \sum_j m_{ij} \tilde{x}_j \tag{19}$$

for weights $m_{ij} = m_{E,ij} - m_{I,ij}$ where E and I represent excitatory and inhibitory synapses respectively, and real valued inputs x. This use of different synapses for excitatory and inhibitory connections is biologically more plausible, but does not affect the calculations otherwise.

The following constraints on the network connections are deduced:

- $m_{I,ij}$ increases if $\tilde{\mathbf{x}}_j$ is high and $\tilde{\mathbf{x}}_i$ is high
- $m_{I,ij}$ decreases if $\tilde{\mathbf{x}}_j$ is high and $\tilde{\mathbf{x}}_i$ is low
- $m_{E,ij}$ increases if $\tilde{\mathbf{x}}_j$ is low and $\tilde{\mathbf{x}}_i$ is high
- $m_{E,ij}$ decreases if $\tilde{\mathbf{x}}_j$ is low and $\tilde{\mathbf{x}}_i$ is low

- in other cases, $m_{I,ij}$ and $m_{E,ij}$ will be stationary,

and used to derive the following linearised model of the network:

$$\frac{dm_{ij}}{dt} = -\alpha \tilde{x_i} \tilde{x_j}, \tag{20}$$

where $\tilde{x_i} = x_i + \sum_j m_{ij} \tilde{x_j}$ and α is a positive constant. So, for inputs $\mathbf{x}$ the network produces outputs $\tilde{\mathbf{x}}$, $(\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^n)$.

Equation 20 can be written in matrix notation as:

$$\frac{dM}{dt} = -\alpha \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T, \tag{21}$$

where the network output is

$$\tilde{\mathbf{x}} = \mathbf{x} + M\tilde{\mathbf{x}} \tag{22}$$

for $M \in \mathbb{R}^{n \times n}$, $M|_{ij} = m_{ij}$, and network transfer function $\Phi \in \mathbb{R}^{n \times n}$ is defined as

$$\Phi \mathbf{x} = (I - M)^{-1} \mathbf{x} = \tilde{\mathbf{x}} \tag{23}$$

$$\Rightarrow \frac{d\Phi^{-1}}{dt} = -\Phi^{-1} \frac{d\Phi}{dt} \Phi^{-1} = -\frac{dM}{dt}, \tag{24}$$

$$\frac{d\Phi}{dt} = -\alpha \Phi^2 \mathbf{x}\mathbf{x}^T \Phi^2, \tag{25}$$

for Φ initially symmetric (and therefore symmetric for all t). Equation 25 forms a matrix Bernoulli equation. Kohonen and Oja (1976) show that stable solutions exist for $\alpha \geq 0$, a result that is extended by reducing the constraints on Φ by Oja (1978).

Kohonen noted the similarity of habituation (see section 9) and novelty filtering, commenting on the functionality of his novelty filter in (Kohonen, 1993) (page 101):

> If this phenomenon [producing a non-zero output only for novel stimuli] were discussed within the context of experimental psychology, it would be termed habituation.

While the analogy is not exact, the essence of the idea is sound – storing inputs and then computing the difference between the current input and the best-matching stored pattern does mean that non-zero output is only seen for novel stimuli, although the mechanism by which the effect is accomplished is certainly not biologically realistic.

This novelty filter detects novelty reliably and has some ability to generalise between similar perceptions, since a perception that is very like another will have very few bits different. This is a very primitive quantification of the amount of novelty, although it assumes that all the bits are equally important, which may not be valid.

3.2 Implementations of the Novelty Filter

In the original description of the novelty filter (Kohonen and Oja, 1976), the network is a fully-connected feedback system, meaning that the computational costs are huge. Ko et al. (1992) shows that the novelty filter can be implemented as an auto-associative network (i.e., a network where the input vector is reproduced at the outputs, as is shown in figure 3) trained using back-propagation of error, see, for example, Bishop (1995). They derive a non-linear transfer function for the hidden layer and show that the resulting network is equivalent to the filter of Kohonen and Oja.

An alternative approach is given by Matsuoka and Kawamoto (1993), who analyse a linear, single-layer network with Hebbian and anti-Hebbian learning rules, and show that, for different learning parameters, the

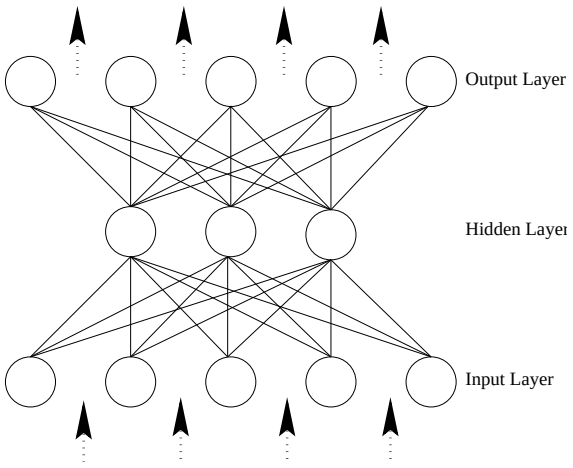

Figure 3: An auto-associative neural network.

network can implement Principal Component Analysis (PCA), orthogonal projection, or novelty filtering of the Kohonen and Oja style.

A different version of the novelty filter implemented as a single-layer network is described in Ko and Jacyna (2000). They propose a continuous, auto-associative perceptron. The update rule for the network weights is converted to a first-order stochastic differential equation, and the paper shows that the probability density function of the weights satisfies the Fokker-Planck equation.

3.3 Variations on the Novelty Filter

Other authors have provided further understanding of the properties of the novelty filter. Aeyels (1990) investigates the convergence properties of the network for more general initial conditions than were used by Kohonen and Oja (1976). He also proposes a modification of the network to allow for the network to forget inputs over time. The network model is then

$$\frac{dM}{dt} = -\alpha\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T - \beta M, \tag{26}$$

with $\beta > 0$ (compare with equation 21). The equivalent transfer function (c.f. equation 25) is then

$$\frac{d\Phi}{dt} = -\alpha\Phi^2\mathbf{x}\mathbf{x}^T\Phi^2 + \beta(\Phi - \Phi^2). \tag{27}$$

The equilibrium solutions of this equation can be analysed as follows. Consider the output of the novelty filter. For a trained input $\mathbf{x}$, the network produces $y(\mathbf{x}) = \mathbf{x}$ at the output. Let a new input be $\mathbf{x}^* = \mathbf{x} + \Delta\mathbf{x}$ for small $\Delta\mathbf{x}$. Then the output of the network is $y(\mathbf{x}^*) = y(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{x}$. The output of the novelty filter is found by subtracting the output of the network from the input, producing

$$\begin{aligned} n(\mathbf{x}^*) &= \mathbf{x}^* - y(\mathbf{x}^*) \\ &= \mathbf{x} + \Delta\mathbf{x} - y(\mathbf{x} + \Delta\mathbf{x}) \\ &= \Delta\mathbf{x}, \end{aligned} \tag{28}$$

which is true if $y(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{x}$.

Elsimary (1996) proposes a new training algorithm for the novelty filter (implemented as an auto-associative network) to ensure that it is insensitive to perturbations in the input patterns. A genetic algorithm

is used to search the weight space of the neural network in order to minimise the network error. The resulting training algorithm is compared to back-propagation for a problem of motor fault detection, and is shown to improve the performance on this task. As was discussed in section 1, novelty detection is often used on fault detection tasks, in both the medical and engineering fields. However, as each researcher investigates their own specific problem, comparisons between the different novelty filters are very difficult. Some applications of the Kohonen and Oja novelty filter are described in the next section.

3.4 Applications

The novelty filter proposed by Kohonen and Oja (1976) has been applied to a number of problems. Kohonen (1993) demonstrates the actions of the filter on a toy problem of characters made from 7×5 blocks of pixels. He also gives a more impressive application, showing the output of the filter when it is presented with radiographic images of the brain. The network is presented with 30 images from normal patients (i.e., patients where no medical problems were diagnosed) for training, and then a number of test images, some normal and some abnormal were presented. The filter highlighted those parts of the images that were not seen in the training set, that is the features that were novel. Since these features were not in the training set of normal images they were presumed to be abnormal, and therefore potential brain tumours. Detailed results are not given, so it is not possible to assess how useful the technique would be in practice.

The novelty filter has also been used for radically different applications. For example, Ardizzone et al. (1990) used it to detect motion in a series of images. The network identifies the trajectory of an object by identifying the novelty in the current image with respect to patterns related to previous positions.

Kohonen and Oja's novelty filter has also been used to detect machine breakdowns and for related engineering tasks. For example, Worden (1997) and Worden et al. (2000) compared the novelty filter to kernel density estimation (described in section 4.2) and to measuring the Mahalanobis distance (see section 2) on a structural fault detection task. Similar results were found for all three methods, although the Mahalanobis distance method was the fastest, requiring little training and very little calculation for the evaluation. The novelty filter is also used by Streifel et al. (1996) to detect shorted turns in turbine-generator rotors.

In a variation on the theme of breakdowns, Chen et al. (1998) use the novelty filter as one of a number of methods of novelty detection (the others are PCA, described in section 2.6, and density estimation (section 4.2)) to detect motorway incidents. All three techniques were found to be successful, although density estimation had a faster response time, but a higher false alarm rate.

3.5 Related Approaches

A similar approach to that of the Kohonen and Oja novelty filter is proposed by Japkowicz et al. (1995). In their scheme, which is part of a model of the hippocampus, an auto-associative network is also used. However, rather than highlighting the novel parts of the input, instead the number of bits that are different between the input and output are counted. If this exceeds some threshold then the current input is considered to be novel. Pomerleau (1992) uses a related approach known as Input Reconstruction Reliability Estimation, which consists of computing the following statistic between an input and its reconstruction using an auto-associative network:

$$\rho(I, R) = \frac{\overline{IR} - \overline{I} \cdot \overline{R}}{\sigma_I \sigma_R}, \tag{29}$$

where $\overline{I}$ is the mean activation value of the input, $\overline{R}$ is the mean activation value of the reconstruction (the output), σ_R is the standard deviation in the activation of the reconstruction and $\overline{IR}$ is the mean of the set formed by the unit-wise product of the input and output images. Pomerleau (1992) uses the value of ρ to evaluate how reliable the output of a back-propagation network trained on the input is. Qualitatively these ideas are similar to the novelty filter.

An interesting claim was made by Daunicht (1991), who stated that the mechanoreceptors found in muscles and tendons (such as those used to rotate the eyeball) provide a basis for auto-association and

novelty detection. The interactions between elastic actuators (i.e., muscles) that act on a rigid body cause this effect, which is shown through the minimisation of the potential energy to find the optimum actuator commands.

The characteristic that links the applications described above is that the class that is wanted is under-represented in the data, so that it is not possible to train a classifier to recognise this class explicitly. That is one application where novelty detection is useful. Kohonen and Oja's novelty filter exhibits many of the hallmarks of novelty detection. The network is trained off-line on positive examples, and then the deviation between the input and the best-matching prototype vector is used to evaluate the novelty of the input. These features will be seen in many of the systems described in the next sections.

4 Novelty Detection Using Supervised Neural Networks

4.1 Introduction

One of the principal uses of artificial neural networks is classification, clustering data into two or more classes. Neural networks can be trained in two ways, *supervised* learning, where each training input vector is paired with a target vector, or desired output and *unsupervised* learning, where the network self-organises to extract patterns from the data without any target information.

This section concentrates on supervised neural networks. Typical examples are the perceptron (Rosenblatt, 1962), and related multi-layer perceptron (McClelland et al., 1986), and the Radial Basis Function (RBF) network (Moody and Darken, 1989). Overviews of these and other networks can be found in any standard text, such as Bishop (1995) or Haykin (1999). These networks adapt the connection weights between layers of neurons in order to approximate a mapping function that models the training data. In the trained network, every input produces an output. For classification tasks this is usually an identifier for the best-matching class. However, there is no guarantee that this best-matching class is a good match, only that it is a better match than the other classes for the set of training data that was used. This is where novelty detection is useful, recognising inputs that were not covered by the training data and that the classifier cannot therefore categorise reliably.

4.2 Kernel Density Estimation

Assuming that the network has been trained well, the main reason why the predictions of the network could be poor is that the dataset that was used to train the network is not representative of the whole set of potential inputs. There are two possible reasons for this:

- there are only a few examples of an important class
- the classification set is incomplete

One interpretation of this is that there is a strong relationship between the reliability of the output of the network and the degree of novelty in the input data. This approach has been taken by Bishop (1994), who evaluated the sum-of-squares error function of the network:

$$\begin{aligned} E &= \sum_{j=1}^{m} \int \left[y_j(\mathbf{x}, \mathbf{w}) - t_j\right]^2 p(\mathbf{x}, t_j) d\mathbf{x} dt_j \\ &= \sum_{j=1}^{m} \int \left[y_j(\mathbf{x}, \mathbf{w}) - \langle t_j | \mathbf{x} \rangle\right]^2 p(\mathbf{x}) d\mathbf{x} + \\ &\quad \sum_{j=1}^{m} \int \left\{\langle t_j^2 | \mathbf{x} \rangle - \langle t_j | \mathbf{x} \rangle^2\right\} p(\mathbf{x}) d\mathbf{x}, \end{aligned} \tag{30}$$

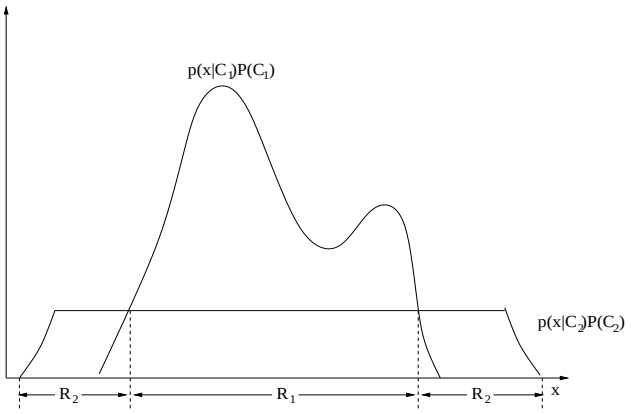

Figure 4: Novelty detection in the Bayesian formalism. The training data is used to estimate $p(\mathbf{x}|C_1)P(C_1)$ using $\hat{p}(\mathbf{x})$, with novel data (class C_2) having a distribution that is assumed constant over some large region. Vectors that are in the regions labelled R_2 are considered to be novel. Adapted from Bishop (1994).

where $p(\mathbf{x}, t_j)$ is the joint probability density function for the data, $j = 1, \ldots, m$ are the output units, $\mathbf{w}$ the weights, $\mathbf{x}$ is the input to the network, t_j the associated target for unit j and y_j the actual output of unit j. The conditional averages of the target data in equation 30 are given by:

$$\langle t_j|\mathbf{x}\rangle \equiv \int t_j p(t_j|\mathbf{x}) dt_j, \tag{31}$$

$$\langle t_j^2|\mathbf{x}\rangle \equiv \int t_j^2 p(t_j|\mathbf{x}) dt_j. \tag{32}$$

Only the first of the two parts of equation 30 is a function of the weights $\mathbf{w}$, so if the network is sufficiently flexible (i.e., the network has enough hidden units), the minimum error E is gained when

$$y_i(\mathbf{x}, \mathbf{w}) = \langle t_j|\mathbf{x}\rangle, \tag{33}$$

which is the regression of the target vector conditioned on the input. The first term of equation 30 is weighted by the density $p(\mathbf{x})$, and so the approximation is most accurate where $p(\mathbf{x})$ is large (i.e., the data is dense).

In general we do not know very much about the density $p(\mathbf{x})$. However, we can generate an estimate $\hat{p}(\mathbf{x})$ from the training data and use this estimate to get a quantitative measure of the degree of novelty for each new input vector. This could be used to put error bars on the outputs, or to reject data where the estimate $\hat{p}(\mathbf{x}) < \rho$ for some threshold ρ, effectively generating a new class of 'novel' data. The distribution of the novel data is generally completely unknown. It can be estimated most simply as being constant over a large region of the input space and zero outside this region to make it possible to normalise the density function, as is shown in figure 4.

This approach was used by Bishop (1994) for data collected from oil pipelines. The training set is first examined by hand to ensure that it has no examples of the novel class, i.e., the class that should be detected. A Parzen window estimator Silverman (1986) with one Gaussian kernel function for each input is then used to model the training set, so that

$$\hat{p}(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}\sigma^d} \sum_{q=1}^{n} \exp\left\{-\frac{|\mathbf{x} - \mathbf{x}^q|^2}{2\sigma^2}\right\}, \tag{34}$$

where $\mathbf{x}^q$ is a data point in the training set and d is the dimensionality of the input space. Any point where the likelihood $\hat{p}(\mathbf{x})$ is below some threshold is considered to be novel.

Tarassenko et al. (1995) followed the same approach, but used a local representation of the data by using the k-means algorithm (Duda and Hart, 1973) to partition the data and then using a different value of the smoothing parameter σ in equation 34 for each data cluster, so that local properties of the data could be taken into account. This approach was tested with reasonable success on various datasets, mostly of fairly small size, including mammograms (Tarassenko et al., 1995) and jet engine data (Nairac et al., 1997, 1999).

A similar approach is taken by Roberts and Tarassenko (1994). A Gaussian mixture model, a method of performing semi-parametric estimation of the probability density function (Tråvén, 1991) is then used to learn a model of the 'normal' data from the training set. The number of mixtures is not defined in advance, with new mixtures being added if the mixture that best represents the data is further from the input that some threshold. In testing, any input that would require a new mixture to be generated is considered to be novel. This technique is capable of continuous learning, as the size of the model grows as more data is provided. The work is applied to a large number of datasets taken from medical problems, such as EEG data for sleep disturbances (Roberts and Tarassenko, 1994), epilepsy (Roberts, 1998) and MRI images of brain tumours (Roberts, 2000).

4.3 Extending the Training Set

Roberts et al. (1994) consider a method of extending the training set so that the neural network can be trained to recognise data from regions that were not included in the original set. Suppose that the training set for some problem spans the region $\mathcal{R} \subset \mathbb{R}^n$. Then

$$p(\text{class}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\text{class}_1)p(\text{class}_1)}{p(\mathbf{x}|\mathcal{R})}, \tag{35}$$

where

$$p(\mathbf{x}|\mathcal{R}) = p(\mathbf{x}|\text{class}_1)p(\text{class}_1) + p(\mathbf{x}|\text{class}_2)p(\text{class}_2). \tag{36}$$

Using Bayes' theorem,

$$\begin{aligned} p(\mathcal{R}|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{R})p(\mathcal{R})}{p(\mathbf{x}|\mathcal{R})p(\mathcal{R}) + p(\mathbf{x}|\mathcal{R}')p(\mathcal{R}')}, \\ &= 1 - p(\mathcal{R}'|\mathbf{x}) \end{aligned} \tag{37}$$

where $\mathcal{R}'$ is the missing class, which is separate from $\mathcal{R}$. The authors then aim to generate data in $\mathcal{R}'$. They do this by generating data and removing any members of it that are in $\mathcal{R}$.

In Parra et al. (1996), the problem of density estimation is considered through minimum mutual information, which is used to factorise a joint probability distribution. A Gaussian upper bound is put on the distribution, and this is used to estimate the density of the probability functions.

Instead of estimating the output density, Tax and Duin (1998) measure the instability of a set of simple classifiers. A number of classifiers are trained on bootstrap samples the same size as the original training set. For new data, the output of all the classifiers is considered. If the data is novel, then the variation in responses from the different classifiers will be large. This approach is applied to three different types of network – a Parzen window estimator, a mixture of Gaussians and a nearest neighbour method. A similar method is employed in Roberts et al. (1996), where a committee of networks (Perrone and Cooper, 1993) initialised at random is used.

4.4 Monitoring the Error Curve

Another method by which novel inputs can be identified in unsupervised learning is by monitoring the error curve, for example the difference between the prediction of the network and the actual data at the next timestep. This was done by Marsland et al. (2001) to select novel inputs as landmarks that were suitable

for mobile robot navigation. A perceptron (Rosenblatt, 1962) was trained on a set of sonar data collected by a robot. The inputs to the network were the sensor perceptions at time t, with the targets being those at time $t+1$. After training, the robot travelled through a set of environments with the perceptron predicting the next set of perceptions. At places where there was novelty, so that the next perception differed from the prediction, a landmark was added to the map that the robot was building. As the output from the perceptron was very noisy, a Kalman filter (Kalman, 1960), see also Maybeck (1990), was used to detect peaks in the curve.

In fact, the Kalman filter can be used to detect changes in the input stream of any process, where these changes indicate some form of novelty. The filter keeps a continuously updated model of the current state of the process (which is assumed to be linear) and inputs that differ from the predicted output by some multiple of the standard deviation can be flagged as novel.

5 The Gated Dipole

5.1 A Description of the Gated Dipole

An animal gets a steady electric shock until it presses a lever. It therefore learns to press the lever. How can a motor response associated with the absence of negative reinforcement become positively reinforcing? This problem was addressed by Grossberg (1972a,b), who proposed a solution known as the gated dipole, a picture of which is given in figure 5.

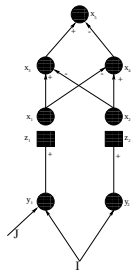

Figure 5: A gated dipole. The value of input J affects which side of the dipole has the higher activity and hence whether x_5 receives excitatory (+) or inhibitory (-) input. Squares represent reservoirs of neurotransmitter. Adapted from Levine and Prueitt (1989).

The gated dipole compares current values of a stimulus with recent past values. It has two channels, which compete. Both receive an arousal stimulus, I, but only one (the left one in figure 5) receives the sensory stimulus J. In the figure, the squares z_1 and z_2 are synapses, with a transmitter that depletes with activity. While the sensory stimulus J is present, the transmitter at z_1 will be depleted more than that at z_2, but the left-hand column will also receive more stimulation and so x_3 will dominate the output. However, when sensory stimulus J is removed, z_2 will have more of the transmitter remaining, and hence x_4 will dominate at the output. Whichever of x_3 and x_4 dominates controls the output of x_5 via the connecting synapse, which is excitatory for x_3 and inhibitory for x_4.

This can be seen more clearly in the equations of the dipole, which are given below in equations 38-46 (g, a_1, a_2 and b are positive constants).

$$\frac{dy_1}{dt} = -gy_1 + I + J \tag{38}$$

$$\frac{dy_2}{dt} = -gy_2 + I \tag{39}$$

$$\frac{dz_1}{dt} = a_1\left(\frac{1}{2} - z_1\right) - a_2 y_1 z_1 \tag{40}$$

$$\frac{dz_2}{dt} = a_1\left(\frac{1}{2} - z_2\right) - a_2 y_2 z_2 \tag{41}$$

$$\frac{dx_1}{dt} = -gx_1 + by_1 z_1 \tag{42}$$

$$\frac{dx_2}{dt} = -gx_2 + by_2 z_2 \tag{43}$$

$$\frac{dx_3}{dt} = -gx_3 + b\left[x_1 - x_2\right]^+, \tag{44}$$

where x^+ denotes $\max(x, 0)$.

$$\frac{dx_4}{dt} = -gx_4 + b\left[x_2 - x_1\right]^+ \tag{45}$$

$$\frac{dx_5}{dt} = -gx_5 + (1 - x_5)\, x_3 - x_5 x_4. \tag{46}$$

On its own, one dipole is not very useful. However, by combining a number of them it is possible to compare stimuli. The combination of gated dipoles is known as a dipole field. For example, Levine and Prueitt (1989, 1992) use a dipole field to model an animal's attention to novelty. They use two dipoles coupled with a reward locus (shown in figure 6). The reward locus provides feedback to each of the competing cues. The first dipole receives some familiar input, while the second dipole receives the test input. The output nodes of the two dipoles ($x_{i,5}$ in figure 6) compete, with the plastic synapses from u favouring cues that have previously won. The only change required to the equations is that equation 46 becomes

$$\frac{dx_{i,5}}{dt} = -gx_{i,5} + (1 - x_{i,5})\left(\alpha u z_{i,5} + x_{i,5}\right) - cx_{i,5}\left(x_{i,4} + \sum_{j \neq i} x_{j,5}\right) \tag{47}$$

(α, c are positive constants), where the new synapses connecting the reward node and the outputs of the dipoles, ($x_{i,5}$, $z_{i,5}$) are controlled by:

$$\frac{dz_{i,5}}{dt} = f_1 z_{i,5} + f_2 u x_{i,5}. \tag{48}$$

In these equations, u is the output of the reward node, which has activity

$$\frac{du}{dt} = gu + r, \tag{49}$$

for some constant r. Note also that each of the dipoles has an inhibitory effect on the others.

5.2 Applications

Levine and Prueitt (1989) used the dipole field to model some data reported by Pribram (1961) that is described in section 8. These experiments demonstrated that monkeys with lesions of the frontal cortex had a preference for novelty. It is these results that Levine et al. wanted to explain through their hypothetical model of the frontal lobes as a dipole field. They found that it is the value of α in equation 47 that controls the gain of signals from the reward locus that is critical. They claim that monkeys with frontal damage have lower values of α and hence the output for novel cues is larger, so that they are favoured. Levine et

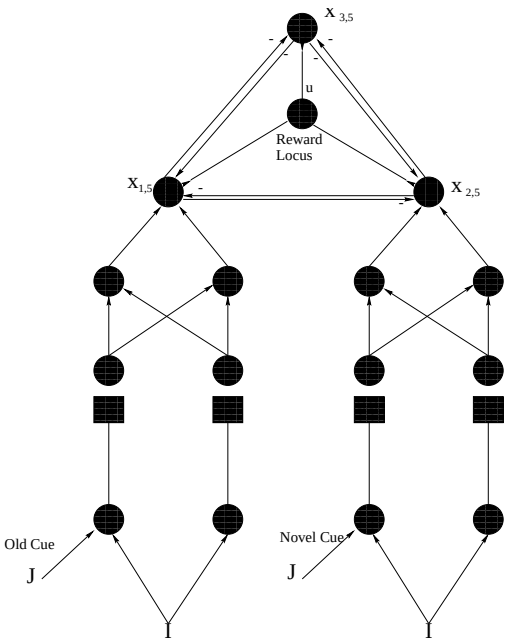

Figure 6: A dipole field. The two dipoles receive different stimuli and compete to see which stimuli to respond to. The reward locus, u, biases the output towards previous winners. Adapted from Levine and Prueitt (1989).

al. claim that their method can be generalised so that further dipoles can be added to represent more cues, which seems reasonable, although they do not provide any evidence of this.

The gated dipole has also been used in a novelty detection system. Öğmen et al. (1992) and Öğmen and Prakash (1997) use the gated dipole in a system that aims to calculate two different types of novelty; spatial novelty and object novelty. Object novelty is used to provide good-bad categorisation using two ART networks. ART is described in section 6.4. The first ART network categorises inputs into object types, while the second categorises the output of the first network into good or bad. Novelty detection in this system is performed by the first ART network. If none of the nodes in the ART network match the input then that input is declared to be novel.

It is in the testing for spatial novelty that the gated dipole is used. The area that the system can see is divided into discrete spatial locations and an array of gated dipoles, one for each spatial location, is used, with the outputs of the dipoles feeding into a winner-takes-all network called the attention spanning module. The system is implemented on a robot arm by Prakash and Öğmen (1998), and the system is used to explore an environment and modify its behaviour according to feedback from the environment.

Systems based on the gated dipole cannot generalise to stimuli that do not have a dipole to represent them. Nor can they scale with the size of the dataset. The amount of transmitter in a dipole does not say how novel a particular stimulus is, but rather how often any novel stimulus has been seen, compared to the non-novel stimulus that the dipole is tuned to recognise.

6 Novelty Detection Methods Based on Self-Organising Networks

This section describes methods that have been used to detect novelty using unsupervised learning algorithms. For novelty detection these algorithms are partially supervised in that, although explicit target vectors are not given, the training set is tailored to ensure that there are no examples of inputs that the network should find to be novel.

6.1 The Self-Organising Map (SOM)

A number of authors have used the Self-Organising Map (SOM) of Kohonen (1982, 1993), one of the best-known and most popular neural networks in the literature. Indeed, a recent survey (Kaski et al., 1998) cited 3343 papers that involved the SOM algorithm in some way, whether analysing it or using it for any of a wide variety of tasks. The algorithm itself, which was inspired by sensory mappings found in the brain, is relatively simple. A lattice of map nodes (neurons) are each connected via adaptable weight vectors linking each node to each element of the input vector. The SOM algorithm performs competitive learning, but instead of just the winning node being adapted, nodes that are close to the winner in the map space (neighbours) are also adapted, although to a lesser extent.

This means that, over time, nodes that are close together in the lattice respond to similar inputs, so that the set of local neighbourhoods self-organise to produce a global ordering. The dimensionality of the map space and the number of nodes in the network are chosen in advance, typically the map is one- or two-dimensional. For a given input, the distance between the input and each of the nodes in the map field is calculated, usually as the Euclidean distance

$$d_j = \sum_{i=0}^{N-1} \left(v_i(t) - w_{ij}(t)\right)^2, \tag{50}$$

where $v_i(t)$ is the input to node i at time t and w_{ij} is the strength of the element of the weight vector between input i and neuron j.

The node with the minimum d_j is selected as the winner, and the weight for that node and its neighbours in the map field are updated using:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)\left(v_i(t) - w_{ij}(t)\right), \tag{51}$$

where j is the index of a node in the neighbourhood, and η is the learning rate ($0 \leq \eta(t) \leq 1$).

In vector form these rules are written

$$d = (\mathbf{v} - \mathbf{w_i})^2 \tag{52}$$

$$\Delta\mathbf{w_i} = \eta(\mathbf{v} - \mathbf{w_i}). \tag{53}$$

There are a number of different ways of initialising the network and of choosing and adapting the neighbourhood size and learning rate. The simplest way of initialising the weights is to give them small random values. Then, to ensure that the network produces a sensible ordering of the weights, a large neighbourhood size and learning rate are used initially, with these variables decreasing over time, for example by an exponential reduction:

$$\eta(t) = e^{\alpha t/\tau}, \tag{54}$$

for constants α and τ, and using a similar equation for the neighbourhood size. This method assumes that the data will be presented in batch for many hundreds of iterations, usually with two different values of α. In the first training regime α is large. This serves to position the nodes in weight space, so that the topology ordering is preserved. Then, during the second phase a smaller value of α is used, so that the learning rate and neighbourhood size are smaller, and the network is fine-tuned to match the input space better. Another approach, which requires that the dataset on which the network will be trained is known in advance, is to use PCA (described in section 2.6) and initialise the nodes in the directions of the first two principal components (or as many principal components as there are dimensions in the map space). In this case the initial training phase with the large learning rate and neighbourhood size is not required.

Although the SOM learning algorithm is very simple, analysis of the network is extremely difficult. There are two important areas for analysis – under what circumstances is the convergence of the network guaranteed, and when will the self-organisation process be, in some sense, optimal. These problems have

been the subject of investigation for a very long time, but only in the case where the map is one-dimensional has a complete analysis been achieved. A survey of the subject is given in Cottrell et al. (1997).

6.2 Measuring the Distance to the Nearest Node

The approach taken by Taylor and MacIntyre (1998) is the simplest. They use a set of data that is known to contain no novelties (i.e., no examples of undesirable features) to train the SOM to recognise a model of normality. Once the SOM is trained, new data is presented to the network and the distance between the node that best matches the new data and any of the nodes that fired when data known to be normal were presented is calculated. If this distance exceeds some threshold then the data used as input is declared to be novel. The aim of the work is to produce machine-monitoring equipment.

A similar kind of idea is used by Höglund et al. (2000) to detect attacks on computers. Their approach, which they term 'anomaly detection', is based on the belief that if the behaviour of a process is consistent it will be concentrated on only a few regions of the feature space. In their case, the processes that they model are the users of a UNIX network, and they are attempting to discover intruders. They compute an anomaly P value by computing the distance to the best-matching unit (BMU) for the current input and then counting the number of BMU distances for the training inputs that are bigger than this distance, and dividing this number by the number of training inputs.

The method proposed by Ypma and Duin (1997) is very different. They investigate how to tell whether two datasets come from the same distribution, as a way of measuring if the second dataset is novel with respect to the first. A number of suitable techniques have been presented in the literature for measuring the quality of the mapping from input space to feature space (see Goodhill and Sejnowski (1997) for a review). Ypma and Duin (1997) use a different measure, a normalised measure of the distance between map units,

$$\begin{aligned} d(x) &= \frac{\|x - m_{p1(x)}\|}{\frac{1}{Q}\sum_{q\in q(x)} \|m_{p1(x)} - m_q\|} + \\ &\left\{ \sum_{n=2}^{k} \alpha_n \left| \min_i \sum_{j=0}^{K_{p_k(x),i}-1} \|m_{I_i(j)} - m_{I_i(j+1)}\| - \right. \right. \\ &\left. \left. \min_l \sum_{j=0}^{K_{q_k(x),l}-1} \|m_{J_l(j)} - m_{J_l(j+1)}\| \right\}, \end{aligned} \tag{55}$$

where m_s is the reference vector of unit s, $I_i(k)$ is the index of the kth unit on the shortest path on the map grid from $I_i(0) = p_1(x)$ (the best match) to $I_{i(K_{p2(x)_1})} = p_2(x)$, the next best match. Similarly, $J_l(j)$ denotes the index of the jth unit on the shortest path l along the map grid from unit $J_l(0) = p_1(x) = q(x)$ to a k-nearest map unit. Q is the cardinality of set q. They also measure the mean-squared error of the mapping. Two problems are investigated using this approach. The first is a variation on the common problem of mechanical fault detection, while the second is the problem of detecting leaks in pipelines.

A way of using the SOM to perform multivariate outlier detection is considered by Muñoz and Muruzábal (1998). They consider two separate techniques:

- graphical methods, plotting the distance between nodes in the map
- measuring the quantisation errors

The first technique is useful for detecting outlying neurons, while the second detects outlying datapoints that project onto neurons that fit into the map well, a feature that may be quite common for high dimensional datasets. The quantisation error is measured by

$$e_k = d\left(\mathbf{x}_k, \mathbf{w}(i^*, j^*)\right), \tag{56}$$

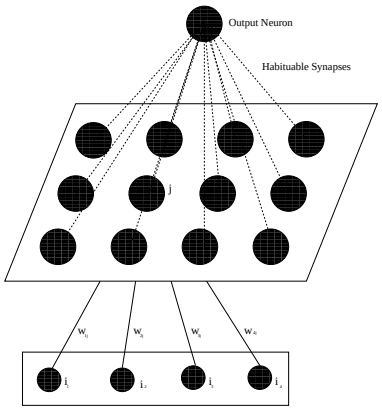

Figure 7: The novelty filter using habituation. The input layer connects to a clustering layer that represents the feature space. The winning neuron (i.e., the one 'closest' to the input) passing its output along a habituable synapse to the output neuron so that the output received from a neuron reduces with the number of times it fires.

where (i^*, j^*) are the coordinates of the winning node and $d(\cdot)$ is the Euclidean distance. They use box plots to decide when an outlier is detected. These approaches are designed to process the results of the mapping for human analysis.

Two methods are used in order to visualise the data for the first technique. One is the Sammon mapping, E, an error calculation that measures the distance between pairs of points, initially in their original space (d_{ij} is the Euclidean distance in that space) and then between the points that represent them in the map space (d'_{ij} is the Euclidean distance in this space) (Sammon, 1969):

$$E = \frac{\sum_j \sum_{i<j} (d_{ij} - d'_{ij})^2}{d_{ij} \sum_j \sum_{i<j} d_{ij}}. \tag{57}$$

This is usually solved by a gradient-descent minimisation technique. The other method is to calculate a matrix of the median-interneuron-distances (MID) over the network.

6.3 Novelty Detection Using Habituation

In the work of Marsland et al. (2002, 2000), the biological phenomenon of habituation is used as part of a novelty detection system. Habituation, a reduction in the strength of a response to a stimulus when the stimulus is seen repeatedly without any ill effects, is a form of novelty filtering in nature and is described in more detail in section 9.

Synapses that are capable of habituation are added to the nodes of several different self-organising networks, to enable the network to act as a novelty detector. Each of the nodes in the map field of the network is linked to an output neuron by an habituable synapse and the winning node of the network propagates its signal along the synapse to the output, while the other nodes do not fire. This synapse then habituates, as, to a lesser extent, do those of the neighbouring neurons. In this way, inputs that are seen frequently are mapped to nodes that have habituated synapses, so that the output of the network for these inputs is very small, but inputs that are seen only rarely causes new nodes to fire, and so the output of the network is very strong.

The approach has been applied to a variety of networks, notably the Self-Organising Map to form the Habituating Self-Organising Map (HSOM), which is shown in figure 7. A new self-organising neural network has also been devised specifically for the task of novelty detection, the Grow When Required (GWR) network (Marsland, 2001; Marsland et al., 2002). This network is capable of adding new nodes into the map field on demand, and yet preserves the topology of the input space.

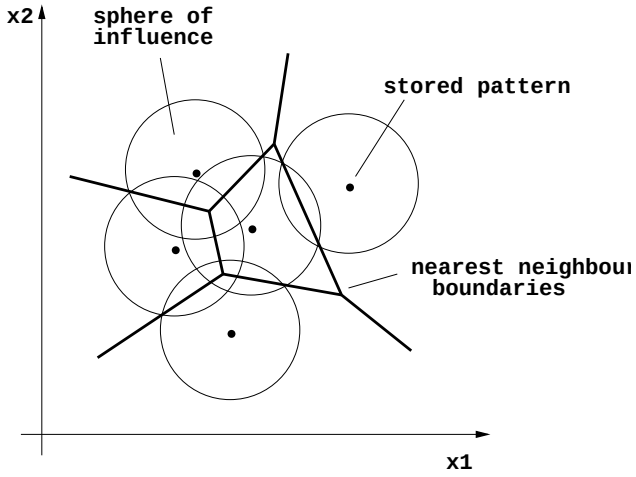

Figure 8: The RCE classifier in two dimensions. A new prototype is created if the input pattern is outside the sphere of influence of all the current stored patterns. Adapted from Kurz (1996).

The GWR network has been used on a number of different tasks, both on-line and batch, ranging from robot inspection of a set of environments using sonar sensors and a camera, to medical diagnosis tasks and machine fault detection.

6.4 The Adaptive Resonance Theory Network

The Adaptive Resonance Theory (ART) network (Carpenter and Grossberg, 1988) attempts to create stable memories from inputs, using match-based learning. The basic ART network performs unsupervised learning. When an input is given to the network, the network searches through the categories currently stored for a match. If a match is found then this category is used to represent the input, if no category is found (so that the strength of response from each of the categories is low) then a new category is added. In itself, this ability to add new nodes whenever none of the current categories represents the data is a form of novelty detection. This approach has been used by a number of researchers, for example, Öğmen et al. (1992) use an ART network to detect object-type novelty, as was described in section 5.2. Caudell and Newman (1993) use an ART network to process a time series, monitoring the creation and usage of the ART categories to see when the time series is stable and when changes occur. A different approach is taken by Lozo (1996), who proposes a match/mismatch detection circuit in a selective attention ART model.

6.5 The Unsupervised Reduced Coulomb Energy Network

The Reduced Coulomb Energy (RCE) network (Reilly et al., 1982) was originally designed for supervised learning of pattern categories. A simplified RCE network for unsupervised learning was introduced by Kurz (1996), who used it to categorise robot sonar readings for navigation. The appearance of the network is shown in figure 8. Input vectors are compared to each of the prototype vectors in the map space, usually using the inner product. The pattern that it is closest in the Euclidean sense to the input is selected as the best match, unless the distance is greater than some threshold (the sphere of influence), in which case the input vector is added into the network as a new prototype. Once added to the network, the prototype vectors do not change.

7 Other Models of Novelty Detection

7.1 Hidden Markov Models

A Hidden Markov Model (HMM) is comprised of a number of states, each with an associated probability distribution, together with the probability of moving between pairs of states (transition probabilities) at each time. The actual state at any time is not visible to the observer (hence the *hidden* part of the name), instead an outcome or observation generated according to the probability distribution of the current state is observed. Further details can be found in Rabiner and Juang (1986). A picture of an HMM is shown in figure 9. HMMs have been found to be very useful in a number of different applications, in particular speech processing (Rabiner, 1989).

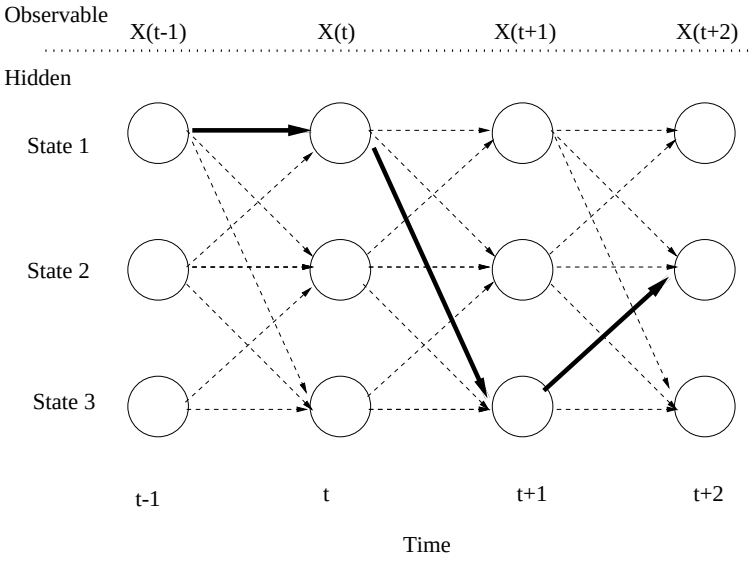

Figure 9: An example of a Hidden Markov Model. Adapted from Smyth (1994b).

As the standard HMM has a predetermined number of states, it is not a useful technique for novelty detection. This has been addressed by Smyth (1994a) to investigate the problem of fault detection in dynamic systems using HMMs. The faults that can occur need to be identified in advance, and states generated for each model. The model also assumes that faults do not happen simultaneously, as this would cause problems with faults not being recognised. The technique is related to the density estimation method that was described in section 4.2, but with the inputs being sequences.

The modification proposed by Smyth (1994b) is to allow extra states to be added while the HMM is being used. Let $w_{\{1,\ldots,m\}}$ be the event that the true system is in one of the states $w_1, \ldots, w_m$, and $p(w_{\{1,\ldots,m\}}|y)$ be the posterior probability that the data is from a known state, given the observation y. Then

$$p(w_i|y) = p_d(w_i|y, w_{\{1,\ldots,m\}})p(w_{\{1,\ldots,m\}}|y),\ 1 \leq i \leq m, \tag{58}$$

where $p_d(\cdot)$ is the posterior probability of being in state i, generated from some discriminative model, and the second part of the product can be calculated using Bayes' rule and the fact that

$$p(w_{m+1}|y) = 1 - p(w_{\{1,\ldots,m\}}|y). \tag{59}$$

The probability of getting a novel state, i.e., a machine fault in the example used by Smyth, can be estimated from the mean time between failures.

7.2 Support Vector Machines

The Support Vector Machine (SVM) is a statistical machine learning technique that performs linear learning by mapping the data into a high dimensional feature space (Vapnik, 1995). The SVM operates by selecting the optimal hyperplane that maximises the minimum distance to the training points that are closest to the hyperplane. This is done in some high dimensional feature space into which the input vectors are mapped using a non-linear mapping called the kernel. For a more detailed description of SVMs, see Burges (1998) or Cristianini and Shawe-Taylor (2000).

SVMs can also be used to describe a dataset (Tax et al., 1999). The aim is to model the 'support' of a data distribution, i.e., a binary valued function that is positive in those parts of the input space where the data lies, and negative otherwise. This means that the SVM can then detect inputs that were not in the training set – novel inputs.

This generates a decision function

$$\text{sign}\left(\mathbf{w}\cdot\Phi(\mathbf{z})+b\right)=\text{sign}\left(\sum_j \alpha_j K(\mathbf{x}_j,\mathbf{z})+b\right), \tag{60}$$

where K is the kernel function (see equation 66), $\Phi(\cdot)$ the mapping into feature space, b the bias, $\mathbf{z}$ the test point, $\mathbf{x}_i$ an element of the training set and $\mathbf{w}$ is the vector

$$\mathbf{w}=\sum_j \alpha_j \Phi(\mathbf{x}_j). \tag{61}$$

A hyperspherical boundary with minimal volume is put around the dataset. This is done by minimising an error function containing the volume of the sphere using Lagrangian multipliers L (R is the radius of the hypersphere):

$$L(R,\mathbf{a},\lambda)=R^2-\sum_i \lambda\left[R^2-(\mathbf{x}_i-\mathbf{a})^2\right],\ \lambda\geq 0, \tag{62}$$

which gives a value of

$$L=\sum_i \alpha_i(\mathbf{x}_i\cdot\mathbf{x}_i)-\sum_{i,j}\alpha_i\alpha_j(\mathbf{x}_i\cdot\mathbf{x}_j), \tag{63}$$

with $\alpha_i\geq 0, \sum_i \alpha_i=1$.

An object $\mathbf{z}$ is considered to be normal if it lies within the boundary of the sphere, i.e.,

$$\begin{aligned}(\mathbf{z}-\mathbf{a})(\mathbf{z}-\mathbf{a})^T &= \left(\mathbf{z}-\sum_i \alpha_i\mathbf{x}_i\right)\left(\mathbf{z}-\sum_i \alpha_i\mathbf{x}_i\right)\\ &= (\mathbf{z}\cdot\mathbf{z})-2\sum_i \alpha_i(\mathbf{z}\cdot\mathbf{x}_i)+\sum_{i,j}\alpha_i\alpha_j(\mathbf{x}_i\cdot\mathbf{x}_j)\\ &\leq R^2.\end{aligned} \tag{64}$$

Further flexibility in the boundary can be gained by replacing the inner products in the above equations by kernel functions $K(x,y)$. Campbell and Bennett (2000) also point out that using slack variables allows certain datapoints to be excluded from the hypersphere, so that the task becomes to minimise the volume of the hypersphere and the number of datapoints outside, i.e.,

$$\min\left[R^2+\lambda\sum_i \xi_i\right] \text{ such that } (\mathbf{x}_i-\mathbf{a})\cdot(\mathbf{x}_i-\mathbf{a})\leq R^2+\xi_i,\ \xi_i\geq 0. \tag{65}$$

In general, this requires quadratic programming to generate a solution. However, Campbell and Bennett (2000) follow Schölkopf et al. (1999) in an alternative approach that significantly reduces the amount of computation required. If the kernel mapping is restricted to be the RBF kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2}, \tag{66}$$

then the data lies on the surface of a hypersphere in the feature space (since $K(\mathbf{x}, \mathbf{x}) = 1$). The hyperplane that separates the surface where the data lies from the region containing no data is generated by constructing the hyperplane that separates all of the datapoints from the origin, but is maximally distant from the origin. Hence, the dual problem is to find

$$\min W(\alpha) = \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \text{ such that } 0 \leq \alpha_i \leq C, \ \sum_{i=1}^{m} \alpha_i = 1. \tag{67}$$

This can be implemented using linear programming techniques.

7.3 The Hopfield Network

The Hopfield network (Hopfield, 1982) is a set of non-linear neurons with binary states, with synaptic connections between every pair of neurons, so that a feedback system is formed. A Hebbian learning rule (Hebb, 1949) is used, so that synapses are strengthened when two neurons fire simultaneously. The network acts as a context addressable memory, storing trained patterns in stable states so that any input that is presented to the network settles into one of the stable states.

The Hopfield network was used by Jagota (1991) to store dictionary words, with the network being used for error correction of text, i.e., as a spelling checker. Jagota (1991) shows experimentally that, although the Hopfield network may occasionally miss an error in the input, even with large dictionaries the network never misclassifies a stored word as a novelty.

One of the key points is that the Hopfield net is not used to retrieve the words, at which task the network does not do well, but merely to state whether or not a particular input is a recognised word or not. This is done by testing whether the state of the network after a word is presented is stable or not. This work has been extended by Bogacz et al. (1999, 2000) in their FamE (Familiarity based on Energy) model. They measure the energy of the network,

$$E(\overline{x}) = -\frac{1}{2} \sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j w_{ij}, \tag{68}$$

where x_i is the value of neuron i, w_{ij} is the strength of the weight vector between neurons i and j, and the sums are over all the N neurons in the network. They suggest that the value of this energy function is lower for stored patterns (of which there are P), being of the order of

$$-N + \mathcal{O}(0, \sqrt{2P}) \tag{69}$$

for stored patterns and

$$\mathcal{O}(0, \sqrt{2P}) \tag{70}$$

for novel patterns that are not correlated with any previous inputs.

They therefore define a threshold halfway between these two values ($-N/2$), and assign any input where the energy of the network is above this threshold to be novel. It is shown that the Hopfield network stores significantly more patterns when the network is not required to reproduce the patterns, merely to classify them. This difference is put at $0.023N^2$ for classification, as opposed to $0.145N$ for retrieval in Bogacz et al. (1999). It is claimed in Bogacz et al. (2000) that the actions of this network are similar to the activities of the perirhinal cortex for familiarity discrimination in animals. This is based on the discovery of so-called

novelty neurons (Brown and Xiang, 1998) that respond maximally to stimuli that have not been seen before, as is described in section 8.

The FamE model is suitable for on-line operation, and has been applied to a robot inspection task (Crook and Hayes, 2001) in work that parallels many of the experiments reported here. A mobile robot travelled parallel to a wall, viewing an 'image gallery' of orange rectangles mounted on the wall using a CCD camera. The image was reduced to a 48×48 binary image, with a 1 at any point signifying the presence of orange, and a 0 the absence of it. This input pattern was presented to a $48 \times 48 (= 2,304)$ node Hopfield network and the energy of the network computed, as described above. If the energy was above the threshold then the current input was set to be novel, otherwise it was considered normal. Experimental observations showed that for a pattern to be found to be normal it had to differ from previously seen patterns by at least 15% (Crook et al., 2002).

7.4 Time to Convergence

Ho and Rouat (1997, 1998) propose a novelty detection method that is based on an integrate-and-fire neuronal model. The usual approach is taken, in that a training set of patterns that are known not to be novel are used to train the network and then test patterns are evaluated with respect to this training set. However, for this technique, it is the time that the network takes to converge when an input is presented that suggests whether or not an input pattern is novel or not.

The network architecture is based on a very simple model of layer IV of the cortex. It consists of a two-dimensional sheet of excitatory and inhibitory neurons with recurrent connections, positioned according to a pseudo-random distribution. Neurons have local connections in a square neighbourhood, with training occurring through Hebbian learning. The state of each neuron is given by

$$S_i(t) = \begin{cases} 0 & \text{if } (t - t_{\text{spike}}) < \rho; \\ \mathcal{H}(U_i(t) - \theta) & \text{otherwise,} \end{cases} \tag{71}$$

where $\mathcal{H}(\cdot)$ is the Heaviside function, $\mathcal{H}(x) = 1$ for $x > 0$ and $\mathcal{H}(x) = 0$ otherwise, and $U_i(t)$ is the control potential,

$$U_i(t) = \sum_i C_{ij} S_j(t+1) + U_i(t-1) + s_i + f_i, \tag{72}$$

for connection strength C_{ij}, input s_i and variable firing frequency function f_i.

The network is applied to 7×5 images of numerical digits, together with noisy versions of the digits, as was done for Kohonen and Oja's novelty filter, and is shown to be superior on this task to a back-propagation network.

7.5 Change Detection

Several authors have investigated change detection, i.e., recognising abrupt changes in signals. This is one part of novelty detection and is related to the methods of monitoring the error curve, such as the Kalman filter, that were described in section 4.4. Abrupt changes may be found when machinery breaks, for example. One approach that has been used in the literature is the Generalised Likelihood Ratio (GLR) test. This is a Neyman-Pearson test (MacDonald, 1997) that decides whether the null hypothesis, that no change occurred in the time between two measurements with unknown probability density functions, is true. The GLR has been implemented as a time-delay neural network and applied to a time-series problem by Fancourt and Principe (2000).

Another approach to change detection is described by Linaker and Niklasson (2000). The purpose of the method is to segment a time series made up of robot sensor data into a number of different phases so that the robot can recognise different parts of an environment, for example, walls, corridors and corners. A similar task was performed by Tani and Nolfi (1999) using a hierarchical mixture of recurrent networks that attempted to predict the next perception based on previous inputs. By way of contrast, Linaker and

Niklasson (2000) propose an adaptive resource allocating vector quantisation (ARAVQ) network. This is based on the idea of finite moving averages of sensor data, encoding the current situation as a function of a bounded interval of past inputs. A change in input is detected when there is a significantly large mismatch between the moving average $\overline{\mathbf{x}}$ of the input and the model vectors $M(t)$, and the situation is stable, i.e., the difference between the values of $\overline{\mathbf{x}}$ for the last n perceptions and the actual last n values is less than some threshold. The approach is shown to be able to differentiate between corridors, corners and rooms using the Khepera simulator, a freeware robot simulator, although the processing is done off-line after all the data has been collected.

7.6 Unusual Approaches to Novelty Detection

A number of other methods have been proposed in the literature for novelty detection. As they do not fall into any of the previous categories, they are described in this section.

7.6.1 Self-Nonself Discrimination

A technique inspired by the action of the immune system is presented in Dasgupta and Forrest (1996). The immune system performs self-nonself discrimination to recognise foreign cells, which are potential infections. This technique has been used for computer virus detection in previous work (D'Haeseleer et al., 1996).

A set of strings are generated that describe the state of the system. From this, a set of detectors are generated that fail to match any of the strings in the set. The match that is measured need only be partial, so that the strings only need to match in r contiguous places, for some value of r. At each timestep a new set of strings that describe the state of the system are generated, and compared to the set of detectors. If a match is found at any time then the state of the system is decreed to be novel. The approach is applied to some data on tool breakdowns and is shown to recognise up to 90% of problems on that dataset.

This technique could be adapted to operate on-line after the initial training stage, with detectors being removed from the test set if they matched inputs that were added at a later date.

7.6.2 The Independent Component Classifier

Linares et al. (1997) describe a feed-forward neural network based on an Independent Component Classifier, with one neuron for each class, and another for the novel class. Each class has an associated prototype. The novelty cell is activated if the current input is significantly outside the space spanned by the current prototypes, i.e., if for input $\mathbf{x}$,

$$\frac{\|\mathbf{x} - \mathbf{x}_{\mathrm{ct}}\|}{\|\mathbf{x}\|} > \rho, \tag{73}$$

where $\mathbf{x}_{\mathrm{ct}}$ is the component of $\mathbf{x}$ inside the prototype space. This approach can be used on-line, since the creation of a new prototype can signify novelty.

7.6.3 A Competitive Learning Tree

A very different approach is described by Martinez (1998), who proposes a neural competitive learning tree that adapts on-line to track time-varying distributions. The current estimated model from the tree is compared to an *a priori* estimate (for example, made from the tree at some previous stage of learning), with a mismatch between the two models signifying novelty.

7.6.4 The Generalised Radial Basis Function Network

A Generalised Radial Basis Function (GRBF) network is used for novelty detection by Albrecht et al. (2000). The GRBF is extended from the RBF network by reverse connections from the output layer back to the central layer are used to make the GRBF self-organise a Bayes classifier (an off-line method). This network is then used as a novelty detector by evaluating the activity of the central layer,

$$A(\mathbf{x}|\theta_c) = \sum_{r=1}^{M} a_r^l(\mathbf{x}|\hat{P}_r, \theta_r), \tag{74}$$

for parameters of each centre $\hat{P}_r$ and θ_r, with a_r^l being defined by

$$a_r^l(\mathbf{x}|\hat{P}_r, \theta_r) = \hat{P}_r \hat{p}(\mathbf{x}|r, \theta_r), \tag{75}$$

where $\hat{p}(\mathbf{x}|r, \theta_r)$ is the multivariate normal distribution. Then, if the activity of the central layer, $A(\mathbf{x}|\theta_c)$, is below some threshold, the current input is considered to be novel.

7.6.5 The Eigenface Algorithm

Hickinbotham and Austin (2000) use a method that is particularly tailored to the problem of detecting faults in aeroplanes. During each flight a frequency of occurrence matrix (FOOM) is generated, with counts of particular stress events. A novelty detection technique is applied to analyse these FOOMs, a technique that it is only possible to perform off-line. This method is related to the Eigenface algorithm (Turk and Pentland, 1991) used to recognise images of faces by computing the first few principal components of the images and then computing the eigenvectors of the covariance matrix that spans the principal component space.

The mean average FOOM, Φ, of the training set $\{\Gamma_n\}$ is calculated, together with the deviation from this mean for each FOOM ($\Psi_n = \Gamma_n - \Phi$) in the training set. Then the eigenvalues λ_k and eigenvectors $\mathbf{v}_k$ of the matrix $\mathbf{L}$ defined by

$$\mathbf{L}_{j,k} = \Psi_j^T \Psi_k \tag{76}$$

are computed. The M eigenvectors with the largest eigenvalues are used to compute the so-called eigen-FOOMs,

$$\mathbf{u}_m = \sum_k \mathbf{v}_m k \Psi_k. \tag{77}$$

A new FOOM is evaluated by computing the deviation of the new FOOM from the mean, and the distance of that to each of the principal components.

8 Biological Novelty Detection

Researchers in biology and psychology have been studying the ability of animals to detect novelty for a long time. This section presents an overview of some of the more salient work so that some of the effects of novelty detection in animals can be seen. Many animals respond to unexpected stimuli with an orienting response that demonstrates that novelty causes fear (O'Keefe and Nadel, 1978). This is followed by an exploration phase, where the animal carefully begins to explore its environment again.

To an animal, items or places that have not been experienced before are novel. The unfamiliar conjunction of familiar objects can also be novel. To O'Keefe and Nadel this implies that memories of items include their context. They describe several sets of experiments by a number of different researchers that have investigated the responses of a variety of animals (fish, rats, gerbils, etc.) to novel stimuli. In one particularly interesting case rats were subjected to three types of novel occurrence:

- introducing a novel item in a familiar environment
  - with the animal engaged in directed activity (competitive)
  - with the animal resting (non-competitive)
- introducing the animal into a novel environment
- spontaneous alteration of the environment

The results of the experiments show an interesting dichotomy between the response of hippocampal animals (that is, animals with hippocampal lesions) and control animals. In general, normal animals appear to be distracted by unexpected features and will explore them, finishing any task that they were already performing later, after they had dealt with the interruption, while hippocampal animals will not explore the novel stimuli. There is evidence that they do recognise that the stimulus is novel, but just ignore it. This suggests that the hippocampus is involved in dealing with novelty and that it controls the extent to which the exploration impulse overcomes fear. So curiosity may have killed the cat, but only if its hippocampus was intact.

The hippocampus has also been shown to be a critical part of the brain network that detects novelty in humans. This was demonstrated by Knight (1996) using electrophysiological recording of scalp event-related potentials (ERPs). A series of target and novel stimuli (tones and finger taps) were embedded in a stream of repetitive background stimuli, and the ERPs recorded throughout the experiment. Experiments on patients with damage to the hippocampus showed that these patients responded less to novel events than control patients. These results are not surprising when the fact that the hippocampus is widely thought to be involved in the processes of memory is considered – any type of novelty detection requires the recall of previously seen stimuli.

One method of testing for novelty detection in humans that is used in the laboratory is to test the *von Restorff* effect. This is the finding that, after presentation of a number of stimuli, recall is better for isolates (i.e., a stimulus that differs from the others along some dimension, such as being a different colour or size) than for non-isolates (von Restorff, 1933; Parker et al., 1998). Evidence from tests like these suggest that novelty detection is important for coding information in memory (Brown, 1996). This makes sense because it may reduce the demands made on long-term memory, if perceptions have been seen before then there is no need to remember them. Further evidence of this is provided by the CHARM model of memory storage and retrieval (Metcalfe, 1993), where the storage of a stimulus in memory depends on an assessment of how similar the current stimulus is to memories already laid down.

Further experiments are described by Pribram (1961, 1992), who tested the attraction of rhesus monkeys with a lesioned frontal cortex (frontal monkeys) to novel stimuli. In the experiments, frontal monkeys and controls were presented with a board that had twelve holes drilled in it. A peanut was placed under one of the holes and that hole was covered with an object. All of the monkeys quickly learnt to uncover the hole and find the peanut. However, when a second (novel) object was introduced to cover another hole, and the peanut placed under the new object, normal monkeys took several trials to learn to lift the new object, rather than the old one where previous experience had shown the object to be. This experiment was repeated using a third object as well as the other two, with the same results. Only after about six objects had been added to the board did normal monkeys associate the reward with the novel cue rather than the one that had previously received the reward. In contrast, frontal monkeys were attracted to the novel stimulus immediately, always went for the new object and were therefore rewarded. There is no report of the experiment in reverse, to see if frontal monkeys would learn to look under an object that was not novel if this was reinforced. It would have been interesting to see if it took frontal monkeys as long to learn to pick the object that was reinforced in the previous trial rather than the novel object.

Another area of the brain that is thought to be important for novelty detection is the perirhinal cortex. Brown and Xiang (1998) demonstrate that this region is instrumental in the judgement of prior occurrence. The authors describe the existence of three types of neuron useful for this task, which they claim are found throughout the anterior inferior temporal cortex (which includes perirhinal cortex): *recency neurons* that fire strongly for perceptions that have been seen recently, whether or not they are familiar, *familiarity neurons* that give information about the relative familiarity of a stimulus, and *novelty neurons* that respond strongly to presentations of novel stimuli.

Animals need to know how to focus on the novel stimuli amongst the huge number of features that are perceived. One way that this is done is by responding less to features that are seen repeatedly without ill effect. In the psychological literature this ability is known as habituation, and is the subject of the next section.

9 Habituation

9.1 What Habituation Is

Habituation is a reversible decrement in behavioural response to a stimulus that is seen repeatedly without any ill effects. It is thought to be one of the simplest examples of plasticity in the brain, and as such has attracted a lot of interest. Habituation can be thought of as a way of defocusing attention from features that are seen often, allowing the animal to concentrate fully on other, potentially more dangerous, experiences. Evidence of habituation can be seen clearly in an animal as simple as the sea snail *Aplysia*. This mollusc has a gill that is withdrawn when its siphon is touched. However, repeated gentle stimulation of the siphon results in an habituated response meaning that the gill is withdrawn less and less, and finally not at all. Repeated series of training show that, while the defensive withdrawal returns over time (dishabituation), further stimulation habituates faster. As well as its simplicity there is another reason why habituation has generated so much interest – it occurs in almost all animals, and affects the behaviour of the animal throughout an experiment. Once an animal has perceived a stimulus several times, it will respond to it less because of habituation. This can clearly have a large impact on the experiment. As Zeaman (1976) puts it,

> "Habituation is like rats and cosmic rays. If you are a psychologist, it is hard to keep them out of your laboratory."

Habituation has been investigated by a large number of researchers, and in a wide variety of different animals, from *Aplysia* (Bailey and Chen, 1983; Byrne and Gingrich, 1989; Castellucci et al., 1978; Greenberg et al., 1987) through to cats (Thompson, 1986) and toads (Ewert and Kehl, 1978; Wang and Arbib, 1991b). There are also books on aspects of habituation, which give a wide overview of the field from the psychological angle (Peeke and Herz, 1973a; Tighe and Leaton, 1976) and the neuronal (Peeke and Herz, 1973b) angle.

A complete definition of habituation was provided by Thompson and Spencer (1966), who defined nine criteria that describe habituation, based on their studies of the phenomenon. Their main aim was to differentiate habituation from other forms of decrement in behaviour, such as fatigue. The criteria are given below:

1. Given that a particular stimulus gives a response, repeated application of the stimulus leads to a decreased response; usually this is a negative exponential function of the number of presentations
2. If the stimulus is withheld, the response recovers over time
3. If repeated series of training and recovery are used, habituation should become more rapid. This is known as potentiation
4. The more rapid the frequency of stimulus presentation, the more rapid the habituation
5. The weaker the stimulus, the more rapid the habituation
6. The effects of habituation training may go beyond zero or the asymptotic response level
7. A stimulus can be generalised to other stimuli
8. Presentation of a non-generalised stimulus leads to dishabituation
9. Repeated application of a dishabituation stimulus leads to a habituation of the amount of dishabituation (that is, if a stimulus is habituated and then dishabituated repeatedly (so that the animal stops responding to it and then responds again), eventually the amount of dishabituation reduces, which means that the animal stays habituated to the stimulus)

9.2 Models of Habituation

A number of authors have described models of the quantitative effects of habituation, as measured with a particular behavioural response. These models can be considered to represent the cellular processes of habituation as they occur in simple organisms.

The first model proposed was that of Stanley (1976). This model is based on the work of Groves and Thompson (1970) and follows the dual-process theory of habituation. The dual-process theory states that the response to a stimulus is controlled by the output of two independent channels, a sensitising process and an habituation process, with the overall behavioural outcome being a summation of the two channels. One of the effects of the sensitisation channel is to enable dishabituation. The two process channels can be seen in figure 10, which shows Stanley's proposed circuit. The output of cells H, X and O are given by equations 78, where I is the external stimulus, labels n, h and s on synapses represent non-plastic, habituating and sensitising synapses respectively, and w_0, w_1, etc. represent the strengths of the synapses.

$$\begin{aligned} H &= w_0 I \\ X &= w_1 H \\ O &= w_2 H + w_3 X. \end{aligned} \tag{78}$$

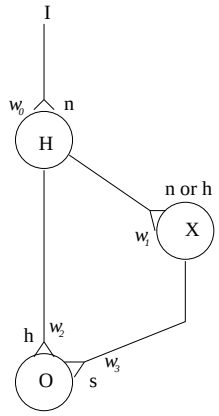

Figure 10: The two process circuit model proposed by Stanley (1976). Cell H receives external input from I, and propagates it to cells X and O via a combination of habituating synapses (h), sensitising synapses (s) and non-plastic synapses (n) that have strengths w_i for $i = 0, .., 3$.

The equation that controls the value of the habituation synapses (labelled h in figure 10) is given in equation 79:

$$\tau \frac{dy(t)}{dt} = \alpha \left[y_0 - y(t) \right] - S(t), \tag{79}$$

where y_0 is the initial value of the weight y, $S(t)$ is the external stimulation and τ and α are time constants governing the rate of habituation and recovery rate respectively. A graph showing the effects of equation 79 is given in figure 11.

Equation 79 can be solved:

$$y(t) e^{\frac{\alpha t}{\tau}} + c = \int \left(\frac{-S(t)}{\tau} + \alpha \frac{y_0}{\tau} \right) e^{\frac{\alpha t}{\tau}} dt. \tag{80}$$

This allows for two different behaviours, depending on whether or not a stimulus S is being applied. Assuming that a constant non-zero stimulus S is applied,

Figure 11: An example of how the synaptic efficacy drops when habituation occurs using Stanley's model (equation 79). In both curves a constant stimulus $S(t) = 1$ is presented, causing the efficacy to fall. The stimulus is reduced to $S(t) = 0$ at time $t = 60$ where the graphs rise again and becomes $S(t) = 1$ again at $t = 100$, causing another drop. The two curves show the effects of varying τ in equation 79. It can be seen that a smaller value of τ causes both the learning and forgetting to occur faster. The other variables were the same for both curves, $\alpha = 1.05$ and $y_0 = 1.0$.

$$y = y_0 - \frac{S}{\alpha}\left(1 - e^{\frac{-\alpha t}{\tau}}\right), \tag{81}$$

and when the stimulus is withdrawn, so that $S = 0$, the solution is

$$y = y_0 - (y_0 - y_1)e^{\frac{-\alpha t}{\tau}}, \tag{82}$$

where the stimulus was withdrawn when the value of y was $y = y_1$. The two behaviours are shown in figure 11. The first behaviour, detailed in equation 81, describes a drop in the efficacy of the synapse, as is seen initially in the figure, while the second behaviour (equation 82) gives the part of the graph where the efficacy recovers to y_0, its original value (between 60 and 100 presentations in figure 11).

Similar models were proposed by Lara and Arbib (1985) and Staddon (1992). Stanley's model (equation 79) only models the short-term effects of habituation, in that once a stimulus has dishabituated, it will not habituate any faster a second time. This is not what is found in biological investigations, where a stimulus that has habituated once habituates faster a second time (potentiation, point 3 in the list of Thompson and Spencer (1966)). The first model incorporating long-term habituation was produced by Wang and Hsu (1990), who used these equations:

$$\tau\frac{dy(t)}{dt} = \alpha z(t)\left[y_0 - y(t)\right] - \beta y(t)S(t), \tag{83}$$

$$\frac{dz(t)}{dt} = \gamma z(t)\left[z(t) - 1\right]S(t). \tag{84}$$

This pair of equations gives an S-shaped curve that displays both short-term and long-term effects of habituation. β controls the speed of habituation and the input $S(t)$ is gated by being multiplied by $y(t)$ instead of the direct input of equation 79. The new variable, $z(t)$, changes slowly compared with $y(t)$ and is used to control the rate of recovery. It has a single point of inflection, above which recovery is rapid, corresponding to short-term memory; below the point of inflection long-term effects dominate.

9.3 The Habituation Hierarchy

Ewert and Kehl (1978) demonstrated the existence of an habituation hierarchy, where some stimuli dishabituate others, but the same is not true the other way round. A pair of stimuli are at the same level in the hierarchy if their effects on each other are symmetrical, that is showing stimulus A followed by stimulus B is equivalent to showing stimulus B followed by stimulus A.

It was suggested by Wang and Ewert (1992) that the cause of the hierarchy is that dishabituation is a return to normal behaviour (the sensitisation channel of the dual-process theory), and cross-talk between the habituation and dishabituation process causes the hierarchy. The paper of Wang and Ewert (1992) is one of a series that attempt to model the neural processes underlying the orienting reflex and prey-catching behaviour in toads (genus *bufo bufo*), and its habituation. The other papers model the tectal relay and anterior thalamus (Wang and Arbib, 1991a,b), the areas that process the image taken from the retina, and the medial pallium (Wang and Arbib, 1992), the analogue in reptiles of the mammalian hippocampus, and the region in which habituation is thought to take place.

The experiments reported in these papers all took the same approach. A toad sat in a cylindrical glass jar and an electrically driven prey dummy was moved at $20°s^{-1}$ in a horizontal plane, 7 cm in front of the toad. When the toad recognised the dummy as prey, the toad followed it by making turning movements to orient itself towards the dummy. The number of orienting responses per minute were measured for the duration of the presentation of the prey dummy in order to measure the amount of habituation. By using a variety of prey dummies with different appearances it was noted that in some cases, after habituating to one dummy, the toad did not respond to another, although this was not true the other way round – if the toad habituated to the second dummy, it still responded to the first. This was taken to demonstrate the existence of the hierarchy.

Habituation has been used in an artificial neural network, too. Stiles and Ghosh (1995, 1997) consider the problem of how dynamic neural networks can be used to recognise temporal patterns in data. Their solution is to include an habituation term on the weights that connect the inputs to the neural network, a multi-layer perceptron or radial basis function network. These weights then have short-term temporal information in them, which affects the dynamics of the network.

10 Conclusions

Novelty detection, recognising that certain elements of a dataset do not fit into the underlying model of the data, is an important problem for learning systems – if data used as input to the system does not correspond with the data in the training set, then the performance of the system will suffer.

The most frequent uses of novelty detection systems are in applications where it is hard to control the training set to ensure that there are examples of every type of input. For example, it may be that one class is under-represented in the data and so a classifier that is trained on the data will not recognise that class. Alternatively, a particular class may be so important that missing any examples of that class is worse than mistakenly classifying data as belonging to that class (i.e., false positives are less important than false negatives). These types of problems are common in machine fault detection and medical diagnosis, as typically there are many more examples of healthy test results than of results that show the disease that should be detected.

A precise definition of novelty detection is hard to arrive at, nor is it possible to suggest what an 'optimal' method of novelty detection would be. For example, for any given method it is hard to provide exact answers to questions such as How different should a novel stimulus be?' and How often must a stimulus be seen before it stops being novel?'. Despite, or perhaps because of, this lack of definition, there have been many novelty detection methods described in the literature. This paper has reviewed those methods.

By far the most common approach is to prepare a training set that is known to contain no examples of the important class (i.e., the examples of the disease in a medical application, which will be the novel class) and then use a learning system of some kind to learn a representation of this dataset. After training, the novelty of test inputs can be evaluated by comparing them with the model that has been acquired from the

training set. This approach is related to the statistical methods of outlier detection that were described in section 2. Examples of these types of novelty detector include Kohonen and Oja's novelty filter and related work (described in section 3), neural network validation (section 4.2) and systems that use the gated dipole (section 5).

The training set is a particularly important part of these novelty detection systems. If the training set contains any examples of the class that should be found to be novel then they may be missed by the novelty filter, which is obviously a problem. Alternatively, if the training set does not include sufficient examples of all the data that should be found to be normal, then inputs that should be recognised by the novelty filter will be highlighted as novel. Ideally, therefore, the filter should be robust to a few examples of the novel class in the training set and it should be possible to add new examples to the trained filter without having to discard that network and train a new network from the beginning with the augmented dataset.

Methods that can deal with one or other of these problems have been devised, for example the novelty filter based on the Hidden Markov Model (see section 7.1) can add new states on-line, so that further training of the filter can be performed after the initial training has been completed. Similarly, the Grow When Required network, which is described in section 6.3, can also be trained continuously, so that examples that were not in the original training set can be added later. The novelty detection capability of the GWR network is based on habituation, described in more detail in section 9, which is one method by which animals learn to recognise stimuli that are seen frequently without ill effects, and can therefore be safely ignored.

The question of which novelty filter to use for any given task depends crucially on the task. For applications where it is easy to generate a a training set of 'normal' data on which to train the filter and where this data will never change, any of the methods reported here would be applicable. If the aim is to detect abrupt changes in, for example, the signature of a piece of operating machinery or some other time series, then techniques such as the Kalman filter (section 4.4) or the Generalised Likelihood Ratio test described in section 7.5 are more suitable.

A number of applications of novelty detection have been identified. By far the most common type of application is where there are insufficient examples of the important class, examples of which are machine fault detection and medical diagnosis, as suggested previously. For these examples the incidence of false positives (i.e., incorrectly identifying inputs as novel) is less important than false negatives, which could have very serious consequences. Novelty detection can also be used for inspection tasks, where a robot (or some other set of sensors) can learn to recognise the inputs that are seen in a normal environment that does not have any failings and then highlight any places where the inputs do not fit the acquired model. Finally, novelty detection can be used to reduce the number of inputs that are seen by other systems, i.e., as a method of preprocessing that enables a learning system to focus its attention onto only those perceptions that have not been seen before, or seen only rarely. For example, a neural network could only learn about data that it has not seen before, or a robot only respond to input stimuli that it has not previously experienced.

As has been demonstrated, novelty detection is often a useful approach for machine learning tasks. This paper has described a variety of different methods of novelty detection and statistical outlier detection, but it is relatively under-represented in the literature and there is still work to be done. An understanding of how the different techniques are related would be useful, as would thorough investigation of how the techniques operate for a variety of different applications, including on-line and off-line learning.

Acknowledgements

This research was completed as part of a PhD at the University of Manchester. The work was performed under the supervision of Dr Ulrich Nehmzow and Dr Jonathan Shapiro, whose help is gratefully acknowledged.

References

Dirk Aeyels. On the dynamic behaviour of the novelty detector and the novelty filter. In B. Bonnard, B. Bride, J.P. Gauthier, and I. Kupka, editors, *Analysis of Controlled Dynamical Systems*, pages 1 – 10, 1990.

S. Albrecht, J. Busch, M. Kloppenburg, F. Metze, and P. Tavan. Generalised radial basis function networks for classification and novelty detection: Self-organisation of optimal Bayesian decision. *Neural Networks*, 13:1075 – 1093, 2000.

E. Ardizzone, A. Chella, S. Gaglio, D. Morreale, and S. Sorbello. The novelty filter approach to detection of motion. In E.R. Caianiello, editor, *Third Italian Workshop on Parallel Architectures and Neural Networks*, pages 301–308, 1990.

Craig Bailey and M.C. Chen. Morphological basis of long-term habituation and sensitization in *aplysia*. *Science*, 220:91–93, 1983.

V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, USA, 3rd edition, 1994.

Christopher M. Bishop. Novelty detection and neural network validation. *IEEE Proceedings on Vision, Image and Signal Processing*, 141(4):217–222, 1994.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, England, 1995. ISBN 0-19-853864-2.

Rafal Bogacz, Malcolm W. Brown, and Christophe Giraud-Carrier. High capacity neural networks for familiarity discrimination. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 773 – 778, 1999.

Rafal Bogacz, Malcolm W. Brown, and Christophe Giraud-Carrier. Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience*, 2000.

M. W. Brown and J.-Z. Xiang. Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Progress in Neurobiology*, 55:149 – 189, 1998.

M.W. Brown. Neuronal responses and recognition memory. *Seminars in the Neurosciences*, 8:23 – 32, 1996.

Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121 – 167, 1998.

John H. Byrne and Kevin J. Gingrich. Mathematical model of cellular and molecular processes contributing to associative and nonassociatiave learning in *aplysia*. In John H. Byrne and William O. Berry, editors, *Neural Models of Plasticity*, chapter 4, pages 58–72. Academic Press, New York, 1989.

Colin Campbell and Kristin P. Bennett. A linear programming approach to novelty detection. In T.K. Leen, T.G. Diettrich, and V. Tresp, editors, *Proceedings of Advances in Neural Information Processing Systems 13 (NIPS'00)*, Cambridge, MA, 2000. MIT Press.

Gail A. Carpenter and Stephen Grossberg. The ART of adaptive pattern recognition by a self–organising neural network. *IEEE Computer*, 21:77 – 88, 1988.

V.F. Castellucci, T.J. Carew, and E.R. Kandel. Cellular analysis of long-term habituation of the gill-withdrawel reflex in *aplysia*. *Science*, 202:1306–1308, 1978.

Thomas P. Caudell and David S. Newman. An adaptive resonance architecture to define normality and detect novelties in time series and databases. In *IEEE World Congress on Neural Networks*, pages 166–176, 1993.

Haibo Chen, Roger D. Boyle, Howard R. Kirby, and Frank O. Montgomery. Indentifying motorway incidents by novelty detection. In *8th World Conference on Transport Research*, 1998.

M. Cottrell, J.C. Fort, and G. Pages. Theoretic aspects of the SOM algorithm. In *Proceedings of Workshop on Self-Organising Maps (WSOM'97)*, pages 246 – 267, 1997.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge University Press, Cambridge, 2000.

Paul Crook and Gillian Hayes. A robot implementation of a biologically inspired method for novelty detection. In *Proceedings of Towards Intelligent Mobile Robots*, 2001.

Paul Crook, Stephen Marsland, Gillian Hayes, and Ulrich Nehmzow. A tale of two filters – on-line novelty detection. In *Proceedings of the International Conference on Robotics and Automation (ICRA'02)*, pages 3894 – 3900, 2002.

Dipanka Dasgupta and Stephanie Forrest. Novelty detection in times series data using ideas from immunology. In *Proceedings of the Fifth International Conference on Intelligent Systems*, 1996.

Wolfgang J. Daunicht. Autoassociation and novelty detection by neuromechanics. *Science*, 253:1289 – 1291, 13 September 1991.

L. Denby and R. D. Martin. Robust estimation of the first order autoregressive parameter. *Journal of the American Statistical Association*, 74:140 – 146, 1979.

Luc Devroye and Gary L. Wise. Detection of abnormal behaviour via nonparametric estimation of the support. *SIAM Journal of Applied Mathematics*, 38(3):480 – 488, 1980.

Patrik D'Haeseleer, Stephanie Forrest, and Paul Helman. An immunological approach to change detection: Algorithms, analysis and implications. In *IEEE Symposium on Security and Privacy*, 1996.

R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, USA, 1973.

Hamed Elsimary. Implementation of neural network and genetic algorithms for novelty filters for fault detection. In *Proceedings of the 39th Midwest Symposium on Circuits and Systems*, pages 1432–1435, 1996.

J.-P. Ewert and W. Kehl. Configurational prey-selection by individual experience in the toad *bufo bufo*. *Journal of Comparative Physiology A*, 126:105–114, 1978.

Craig L. Fancourt and Jose C. Principe. On the use of neural networks in the generalised likelihood ratio test for detecting abrupt changes in signals. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'00)*, volume II, pages 243 – 248, 2000.

Geoffrey J. Goodhill and Terrence J. Sejnowski. A unifying objective function for topographic mappings. *Neural Computation*, 9:1291–1304, 1997.

S. Greenberg, V. Castellucci, H. Bayley, and J. Schwartz. A molecular mechanism for long-term sensitisation in *aplysia*. *Nature*, 329:62–65, 1987.

Stephen Grossberg. A neural theory of punishment and avoidance. I. Qualitative theory. *Mathematical Biosciences*, 15:39–67, 1972a.

Stephen Grossberg. A neural theory of punishment and avoidance. II. Quantitative theory. *Mathematical Biosciences*, 15:253–285, 1972b.

P.M. Groves and R.F. Thompson. Habituation: A dual-process theory. *Psychological Review*, 77(5):419–450, 1970.

E.J. Gumbel. *Statistics of Extremes*. Columbia University Press, New York, USA, 1958.

J. Hájek and Z. Šidák. *Theory of Rank Tests*. Academic Press, New York, USA, 1967.

Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, New Jersey, USA, 2nd edition, 1999.

D.O. Hebb. *The Organisation of Behaviour*. Wiley, New York, 1949.

Simon J. Hickinbotham and James Austin. Neural networks for novelty detection in airframe strain data. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'00)*, volume VI, pages 375 – 380, 2000.

Tuong Vinh Ho and Jean Rouat. A novelty detector using a network of integrate and fire neurons. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN'97)*, pages 103 – 108, 1997.

Tuong Vinh Ho and Jean Rouat. Novelty detection based on relaxation time of a network of integrate–and–fire neurons. In *Proceedings of the 2nd IEEE World Congress on Computational Intelligence (WCCI'98)*, pages 1524–1529, 1998.

D.C. Hoaglin, F. Mosteller, and J.W. Tukey, editors. *Understanding Robust and Exploratory Data Analysis*. John Wiley, New York, 1983.

Albert J. Höglund, Kimmo Hätönen, and Antti Sorvari. A computer host-based user anomaly detection system using the self-organising map. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'00)*, volume V, pages 411 – 416, 2000.

J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, volume 79, pages 2554–2558, USA, 1982.

Peter J. Huber. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, USA, 1981.

Arun Jagota. Novelty detection on a very large number of memories stored in a Hopfield-style network. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'91)*, 1991.

Natalie Japkowicz, Catherine Myers, and Mark Gluck. A novelty detection approach to classification. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 518 – 523, 1995.

I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, Berlin, Germany, 1986.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:34 – 45, March 1960.

S. Kaski, J. Kangas, and T. Kohonen. Bibliography of self-organising map (SOM) papers: 1981 – 1997. *Neural Computing Surveys*, 1:102 – 350, 1998.

R. T. Knight. Contribution of human hippocampal region to novelty detection. *Nature*, 383:256 – 259, 1996.

Hanseok Ko, Robert Baran, and Mohammed Arozullah. Neural network based novelty filtering for signal detection enhancement. In *Proceedings of the 35th Midwest Symposium on Circuits and Systems*, pages 252–255, 1992.

Hanseok Ko and Garry M. Jacyna. Dynamical behaviour of autoassociative memory performing novelty filtering for signal enhancement. *IEEE Transactions on Neural Networks*, 11(5):1152 – 1161, 2000.

Teuvo Kohonen. Self-organised formation of topologically correct feature maps. *Biological Cybernetics*, 43: 59–69, 1982.

Teuvo Kohonen. *Self-Organization and Associative Memory, 3rd ed.* Springer, Berlin, 1993.

Teuvo Kohonen and E. Oja. Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biological Cybernetics*, 25:85–95, 1976.

Andreas Kurz. Constructing maps for mobile robot navigation based on ultrasonic range data. *IEEE Transactions on Systems, Man and Cybernetics — Part B: Cybernetics*, 26(2):233–242, 1996.

Rolando Lara and M.A. Arbib. A model of the neural mechanisms responsible for pattern recognition and stimulus specific habituation in toads. *Biological Cybernetics*, 51:223–237, 1985.

Daniel S. Levine and Paul S. Prueitt. Modelling some effects of frontal lobe damage – novelty and perseveration. *Neural Networks*, 2:103 – 116, 1989.

Daniel S. Levine and Paul S. Prueitt. Simulations of conditioned perseveration and novelty preference from frontal lobe damage. In Michael L. Commons, Stephen Grossberg, and John E.R. Staddon, editors, *Neural Network Models of Conditioning and Action*, chapter 5, pages 123 – 147. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

Fredrik Linaker and Lars Niklasson. Times series segmentation using an adaptive resource allocating vector quantisation network based on change detection. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'00)*, volume VI, pages 323 – 328, 2000.

Georges Linares, Pascal Nocera, and Henri Meloni. Model breaking detection using independent component classifier. In *Proceedings of Fifth International Conference on Artificial Neural Networks (ICANN'97)*, pages 559 – 563, 1997.

Peter Lozo. Neural circuit for match/mismatch, familiarity/novelty and synchronization detection in SAART networks. In *International Symposium on Signal Processing and its Applications*, pages 549–552, 1996.

R.R. MacDonald. On statistical testing in psychology. *British Journal of Psychology*, 88:333 – 347, 1997.

Stephen Marsland. *On-line Novelty Detection Through Self-Organisation, With Application to Inspection Robotics.* PhD thesis, Department of Computer Science, University of Manchester, 2001.

Stephen Marsland, Ulrich Nehmzow, and Tom Duckett. Learning to select distinctive landmarks for mobile robot navigation. *Robotics and Autonomous Systems*, 37:241 – 260, 2001.

Stephen Marsland, Ulrich Nehmzow, and Jonathan Shapiro. Novelty detection on a mobile robot using habituation. In *From Animals to Animats: Proceedings of the 6th International Conference on Simulation of Adaptive Behaviour (SAB'00)*, pages 189 – 198. MIT Press, 2000.

Stephen Marsland, Jonathan Shapiro, and Ulrich Nehmzow. A self-organising network that grows when required. *Neural Networks*, 2002.

Dominique Martinez. Neural tree density estimation for novelty detection. *IEEE Transactions on Neural Networks*, 9(2):330 – 338, 1998.

Kiyotoshi Matsuoka and Mitsuri Kawamoto. A self-organising neural network for principal component analysis, orthogonal projection and novelty filtering. In *World Congress on Neural Networks (WCNN'93)*, volume II, pages 501 – 504, 1993.

Peter S. Maybeck. The Kalman filter: An introduction to concepts. In I. Cox and P. Wilfong, editors, *AI-based Mobile Robots: Case Studies of Successful Robot Systems*, pages 193–204. Springer, Berlin, 1990.

James L. McClelland, David E. Rumelhart, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models.* MIT Press, Cambridge, MA, 1986.

Janet Metcalfe. Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review*, 100(1):3 – 22, 1993.

John Moody and Christian J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281 – 294, 1989.

Alberto Muñoz and Jorge Muruzábal. Self-organising maps for outlier detection. *Neurocomputing*, 18:33 – 60, 1998.

Alexandre Nairac, Timothy Corbett-Clark, Ruth Ripley, Neil Townsend, and Lionel Tarassenko. Choosing an appropriate model for novelty detection. In *Proceedings of the Fifth Conference on Artificial Neural Networks (ICANN'97)*, pages 442 – 447, 1997.

Alexandre Nairac, Neil Townsend, Roy Carr, Steve King, Peter Cowley, and Lionel Tarassenko. A system for the analysis of jet system vibration data. *Integrated Computer-Aided Engineering*, 6(1):53 – 65, 1999.

H. Öğmen, R.V. Prakash, and M. Moussa. Some neural correlates of sensorial and cognitive control of behaviour. In *Proceedings of the SPIE, Science of Neural Networks*, volume 1710, pages 177 – 188, 1992.

Haluk Öğmen and Ramkrishna Prakash. A developmental perspective to neural models of intelligence and learning. In Daniel S. Levine and Wesley R. Elsberry, editors, *Optimality in Biological and Artificial Networks?*, chapter 18, pages 363 – 395. Lawrence Erlbaum Associates, Hillsdale, NJ, 1997.

Erkki Oja. S-orthogonal projection operators as asymptotic solutions of a class of matrix differential equations. *SIAM Journal of Mathematical Analysis*, 9(5):848 – 854, October 1978.

John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map.* Oxford University Press, Oxford, England, 1978.

Amanda Parker, Edward Wilding, and Colin Akerman. The von Restorff effect in visual object recognition in humans and monkeys: The role of frontal/perirhinal interaction. *Journal of Cognitive Neuroscience*, 10 (6):691 – 703, 1998.

Lucas Parra, Gustavo Deco, and Stefan Miesbach. Statistical independence and novelty detection with information preserving non-linear maps. *Neural Computation*, 8(2):260 – 269, 1996.

Harman V.S. Peeke and Michael J. Herz, editors. *Habituation*, volume 1: Behavioural Studies. Academic Press, New York, 1973a.

Harman V.S. Peeke and Michael J. Herz, editors. *Habituation*, volume 2: Physiological Substrates. Academic Press, New York, 1973b.

Roger Penrose. A generalized inverse for matrices. In *Proceedings of the Cambridge Philosophical Society*, volume 52, pages 406–413, 1955.

Michael P. Perrone and Leon N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In R.J. Mammone, editor, *Neural Networks for Speech and Image Processing*, chapter 10. Chapman–Hall, New York, USA, 1993.

Dean A. Pomerleau. Input reconstruction reliability esimation. In Stephen Josè Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems 5 (NIPS'92)*, pages 279 – 286, 1992.

Ramkrishna Prakash and Haluk Öğmen. Self-organisation via active exploration: Hardware implementation of a neural robot. *Robotica*, 16:127 – 141, 1998.

Karl H. Pribram. A further experimental analysis of the behavioural deficit that follows injury to the primate frontal cortex. *Experimental Neurology*, 3:432–466, 1961.

Karl H. Pribram. Familiarity and novelty: The contributions of the limbric forebrain to valuation and the processing of relevance. In Daniel S. Levine and Samuel J. Leven, editors, *Motivation, Emotion and Goal Direction in Neural Networks*, chapter 10, pages 337 – 365. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 285, 1989.

Lawrence R. Rabiner and B.H. Juang. An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, pages 4–16, January 1986.

Douglas L. Reilly, Leon N. Cooper, and Charles Erlbaum. A neural model for category learning. *Biological Cybernetics*, 45:35 – 41, 1982.

Stephen Roberts. Novelty detection using extreme value statistics. *IEE Proceedings on Vision, Image and Signal Processing*, 146(3):124 – 129, 1998.

Stephen Roberts, William Penny, and David Pillot. Novelty, confidence and errors in connectionist systems. *Proceedings of the IEE Colloquium on Intelligent Systems and Fault Detection*, 261(10):1 – 10, 1996.

Stephen Roberts and Lionel Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6:270 – 284, 1994.

Stephen Roberts, Lionel Tarassenko, James Pardey, and David Siegwart. A validation index for artificial neural networks. In *Proceedings of the 1st International Conference and Expert Systems in Medicine and Healthcare (NNESMED'94)*, pages 23 – 30, 1994.

Stephen J. Roberts. Extreme value statistics for novelty detection in biomedical data processing. In *Proceedings of the International Conference on Advances in Medical Signal and Information Processing (MEDSIP'00)*, 2000.

F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan, Washington, DC, 1962.

P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, USA, 1987.

J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5): 401 – 409, 1969.

Bernhard Schölkopf, John C. Platt, John Shawe–Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high–dimensional distribution. Technical Report MSR–TR–99–87, Microsoft Research, 1999.

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1986.

Padhraic Smyth. Hidden Markov models for fault detection in dynamic systems. *Pattern Recognition*, 27 (1):149 – 164, 1994a.

Padhraic Smyth. Markov monitoring with unknown states. *IEEE Journal on Selected Areas in Communications*, 12(9):1600 – 1612, 1994b.

J.E.R. Staddon. A note on rate-sensitive habituation. In *From Animals to Animats, Proceedings of the Second International Conference on Adaptive Behaviour (SAB'92)*, pages 203 – 207, 1992.

James C. Stanley. Computer simulation of a model of habituation. *Nature*, 261:146–148, 1976.

Bryan Stiles and Joydeep Ghosh. A habituation based neural network for spatio-temporal classification. In F. Girosi, J. Makhoul, El Manolakos, and E. Wilson, editors, *Proceedings of the Fifth IEEE Workshop on Neural Networks for Signal Processing*, pages 135 – 144, 1995.

Bryan Stiles and Joydeep Ghosh. A habituation based neural network for spatio-temporal classification. *Neurocomputing*, 15(3):273 – 307, 1997.

Robert J. Streifel, R.J. Marks II, M.A. El-Sharkawi, and I. Kerszenbaum. Detection of shorted-turns in the field winding of turbine-generator rotors using novelty detectors – development and field test. *IEEE Transactions on Energy Conversion*, 11(2):312 – 317, 1996.

J. Tani and S. Nolfi. Learning to perceive the world as articulated: an approach for heirarchical learning in sensory-motor systems. *Neural Networks*, 12:1131–1141, 1999.

L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th IEE International Conference on Artificial Neural Networks (ICANN'95)*, pages 442 – 447, 1995.

David M.J. Tax and Robert P.W. Duin. Outlier detection using classifier instability. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 593–601, Berlin, 1998. Springer.

David M.J. Tax, Alexander Ypma, and Robert P.W. Duin. Support vector data description applied to machine vibration analysis. In M. Boasson, J.A. Kaandorp, J.F.M. Tonino, and M.G. Vosselman, editors, *Annual Conference of the Advanced School for Computing and Imaging (ASCI'99)*, pages 398 – 405, 1999.

Odin Taylor and John MacIntyre. Adaptive local fusion systems for novelty detection and diagnostics in condition monitoring. In *SPIE International Symposium on Aerospace/Defense Sensing*, 1998.

R.F. Thompson. The neurobiology of learning and memory. *Science*, 233:941–947, 1986.

R.F. Thompson and W.A. Spencer. Habituation: A model phenomenon for the study of neuronal substrates of behaviour. *Psychological Review*, 73(1):16–43, 1966.

Thomas J. Tighe and Robert N. Leaton, editors. *Habituation: Perspectives from Child Development, Animal Behaviour, and Neurophysiology.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1976.

Hans G.C. Tråvén. A neural network approach to statistical pattern classification by "semiparametric" estimation of probability density functions. *IEEE Transactions on Neural Networks*, 2(3):366 – 377, 1991.

M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71 – 86, 1991.

Vladimir Vapnik. *The Nature of Statistical Learning.* Springer-Verlag, Berlin, 1995.

Hedwig von Restorff. Analyse von vorgangen in spurenfeld (an analysis of the processes in the trace field). *Psychologice Forschung*, 18:299 – 342, 1933.

DeLiang Wang and Michael A. Arbib. Hierarchical dishabituation of visual discrimination in toads. In *From Animals to Animats 1: Proceedings of First International Conference on the Simulation of Adaptive Behaviour (SAB'91)*, pages 77–88, 1991a.

DeLiang Wang and Michael A. Arbib. How does the toad's visual system discriminate different worm-like stimuli? *Biological Cybernetics*, 64:251–261, 1991b.

DeLiang Wang and Michael A. Arbib. Modelling the dishabituation hierarchy: The role of the primordial hippocampus. *Biological Cybernetics*, 76:535–544, 1992.

DeLiang Wang and Jorg-Peter Ewert. Configuration pattern discrimination responsible for dishabituation in common toads *bufo bufo (l.)*: Behavioural tests of the predictions of a neural model. *Journal of Comparative Physiology A*, 170:317–325, 1992.

DeLiang Wang and Chochun Hsu. SLONN: A simulation language for modelling of neural networks. *Simulation*, 55:69–83, 1990.

K. Worden. Structural fault detection using a novelty measure. *Journal of Sound and Vibration*, 201(1):85 – 101, 1997.

K. Worden, S.G. Pierce, G. Manson, W.R. Philp, W.J. Staszewski, and B. Culshaw. Detection of defects in composite plates using lamp waves and novelty detection. *International Journal of Systems Science*, 31 (11):1397 – 1409, 2000.

Alexander Ypma and Robert P. Duin. Novelty detection using self-organizing maps. In *Proceedings of International Conference on Neural Information Processing and Intelligent Information Systems (ICONIP'97)*, pages 1322 – 1325, 1997.

David Zeaman. The ubiquity of novelty – familiarity (habituation?) effects. In Thomas J. Tighe and Robert N. Leaton, editors, *Habituation: Perspectives from Child Development, Animal Behaviour, and Neurophysiology*, chapter 9. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1976.