Novelty Detection: A Review
Part 1: Statistical Approaches

Markos Markou and Sameer Singh
PANN Research, Department of Computer Science
University of Exeter, Exeter EX4 4PT, UK
{m.markou, s.singh}@ex.ac.uk

Abstract
Novelty detection is the identification of new or unknown data or signal that a machine learning system is not aware of during training. Novelty detection is one of the fundamental requirements of a good classification or identification system since sometimes the test data contains information about objects that were not known at the time of training the model. In this paper we provide state-of-the-art review in the area of novelty detection based on statistical approaches. The second part paper details novelty detection using neural networks. As discussed, there are a multitude of applications where novelty detection is extremely important including signal processing, computer vision, pattern recognition, data mining, and robotics.

1. Introduction
Detecting novel events is an important ability of any signal classification scheme. Given the fact that we can never train a machine learning system on all possible object classes whose data the system is likely to encounter, it becomes important that it is able to differentiate between known and unknown object information during testing. It has been realised in practice by several studies that the novelty detection is an extremely challenging task. It is for this reason that there exist several models of novelty detection that have been shown to perform well on different data. It is clearly evident that there is no single best model for novelty detection and the success depends not only on the type of method used but also statistical properties of data handled.

Several applications require the classifier to act as a detector rather as a classifier, that is, the requirement is to detect whether an input is part of the data that the classifier was trained on or it is in fact unknown. This technique is useful in applications such as fault detection (Tarassenko, 1995; Dasgupta and Forrest, 1996; Dasgupta and Nino, 2000; King *et al.*, 2002), radar target detection (Carpenter *et. al.*, 1997), detection of masses in mammograms (Tarassenko, 1999), hand written digit recognition (Tax and Duin, 1998), Internet and e-commerce (Manikopoulos and Papavassiliou, 2002), statistical process control (Guh *et al.*, 1999), and several others. Recently, there has been an increased interest in novelty detection as a number of research articles have appeared on autonomous systems based on adaptive machine learning. However, only a very few surveys have appeared, e.g. Odin and Addison (2000). Much of earlier work and interest in novelty detection sprung from the study of control systems. High integrity systems could not use the traditional classification method for a number of reasons; abnormalities are very rare or there may be no data that describes the fault conditions. Novelty detection offered a solution to this problem by modelling normal data and using a distance measure and a threshold for determining abnormality. In recent years novelty detection has been used in a number of other applications especially signal processing and image analysis (e.g. biometrics). In these applications the problem becomes more complicated with multiple classes, high dimensionality, noisy features and quite often not enough samples. As such, novelty detection methods have tried to keep up with these problems to offer solutions that can be used in the real world. In this paper we review some of the currently used methods on novelty detection using statistical approaches.

There are several important issues related to novelty detection. We can summarise them in terms of the following principles.

a) *Principle of robustness and trade-off:* A novelty detection method must be capable of robust performance on test data that maximises the exclusion of novel samples while minimising the exclusion of known samples. This trade-off should be, to a limited extent, predictable and under experimental control.
b) *Principle of uniform data scaling:* In order to assist novelty detection, it should be possible that all test data and training data after normalisation lie within the same range (Singh and Markou, 2003).
c) *Principle of parameter minimisation:* A novelty detection method should aim to minimise the number of parameters that are user set.
d) *Principle of generalisation:* The system should be able to generalise without confusing generalised information as novel (Tax and Duin, 1998).
e) *Principle of independence:* The novelty detection method should be independent of the number of features, and classes available and it should show reasonable performance in the context of imbalanced dataset, low number of samples, and noise.
f) *Principle of adaptability:* A system that recognises novel samples during test should be able to use this information for retraining (Saunders and Gero, 2000).
g) *Principle of computational complexity:* A number of novelty detection applications are on-line and therefore the computational complexity of a novelty detection mechanism should be as less as possible.

In this survey, we study a number of approaches to novelty detection and remark on how well these studies address the above principles. Each approach has a number of different methods and we detail of the important studies in these areas.

2. Statistical Approaches

Statistical approaches are mostly based on modelling data based on its statistical properties and using this information to estimate whether a test samples comes from the same distribution or not. The techniques used vary in terms of their complexity (Odin and Addison, 2000). The simplest approach can be based on constructing a density function for data of a known class, and then assuming that data is normal computing the probability of a test sample of belonging to that class. The probability estimate can be thresholded to signal novelty. Another simple model can simply find the distance of the sample from a class mean and threshold on the basis of how many standard deviations away the sample is (Manson *et al.*, 2000, 2001). The distance measure itself can be Mahalanobis or some other probabilistic distance (Webb, 1999). Manson *et al*. (2002) also discuss the choice of features based on their ability for novelty detection. In such a scheme a novelty index is used to rank features on their ability for detecting novelty. Another simple statistical scheme for outlier rejection is based on the use of box-plots (Laurikkala, 2000). The box plot is a well-known display of the five-number summary (lower extreme, lower quartile, median, upper quartile, upper extreme). Box plots are most suitable for exploring both symmetric and skewed quantitative univariate data, but they can also identify infrequent values from categorical data. The data (univariate) is sorted in ascending order and the outliers are ranked according to the frequencies of univariate outlier values. Samples with the highest frequencies are discarded. A predetermined percentage of the worst examples can be discarded. Knorr *et al*. (2000) suggest that an object O in a dataset T is a Distance-Based (DB) outlier if at least fraction p of the objects in T lie at a distance greater than D from O. More specifically, based on a standard multidimensional indexing structure, they execute a range search with radius D for each object O. Once there are at least M neighbours in the D-neighbourhood they stop the search and declare O a non-outlier otherwise O is rejected as an outlier. The technique is based in other words on a Nearest Neighbour scheme such as Hellman (1970) and Fumera *et. al.* (2000). The main contribution of this study is that the authors present a solution for fast indexing and search in large multidimensional databases. Several advanced statistical modelling techniques also exist for novelty detection. For example, one can use mixture

models for modelling complex data distributions or using a hidden markov model for novelty detection as discussed later.

Two main approaches exist to the estimation of the probability density function, parametric and non-parametric methods (Desforges *et al.,* 1998). The parametric approach assumes that the data comes from a family of known distributions, such as the normal distribution and certain parameters are calculated to fit this distribution. However, in most real world situations the underlying distribution of the data is not known therefore such techniques have little practical importance. In non-parametric methods the overall form of the density function is derived from the data as well as the parameters of the model. As a result non-parametric methods give greater flexibility in general systems. One of the most intuitive and widely used non-parametric technique is histogram analysis. However, the shape of the density estimate can vary significantly, depending on the choice of the origin, which is often purely arbitrary. Things get even worse when the density estimation of multivariate data is required. A better way of estimating density functions is kernel methods. They are similar to histograms in that they are built up of a number of individual kernels centred on the sampled data points. A parameter h determines the width of the kernel and how smooth the estimation becomes. The kernel function is usually a symmetric probability density function, being non-negative over its domain and should integrate to unity over the defined range. The value of the density at some arbitrary value x is dependent on the distance between the observed data and the shape of the kernel. Clearly, the smoothing parameter h plays a central role on the estimation of the density. Choosing a global value h might result in a density function that does not adequately describe the data, particularly in regions of low data density (various techniques may be employed to locally adjust h).

Parametric methods for estimating the probability density function have sometimes limited use since they require extensive *a priori* knowledge of the problem. Non-parametric statistical approaches make no assumption on the form of data distribution and are therefore more flexible (though more computationally expensive). Density estimation in these cases can be performed using either nearest neighbour methods or Parzen window method. Once more, a probability estimate of the test sample belonging to the distribution can be obtained which can be thresholded. These approaches have limitations with regards to the choice of parameters (neighbours used, smoothing parameters), and difficulty in tackling noise in data. We discuss the parametric and non-parametric approaches in detail in sections 2.1 and 2.2 respectively.

2.1 *Parametric Approaches*

Parametric approaches make an assumption that data distributions are Gaussian in nature and they can be modelled statistically based on data means and covariance. A number of studies have theoretically investigated novelty detection for such data. Of particular importance is the trade-off between the recognition rate and the proportion of data rejected (Hansen *et al*., 1997). The error rate and the reject rate are commonly used to describe the performance level of a pattern recognition system. Because of uncertainty and noise inherent in any pattern recognition task, errors are generally unavoidable. The option to reject is introduced to safeguard against excessive misclassification. However, the trade-off between the errors and rejects is seldom one to one. Whenever the reject option is exercised, some outliers of known classes are also rejected. Chow(1970) studied the trade-off in detail to find an optimal threshold for rejection. It is obvious that a recognition rule is optimum if for a given error rate (error probability) it minimizes the reject rate (reject probability). The author showed that the optimum rule is to reject a pattern if the maximum of the *a posteriori* probabilities is less than some threshold. The error and reject trade-off were derived for the Bayes optimum recognition system with an option to reject.

Hansen *et al*. (1997) extended the work of Chow by introducing the role of classifier confidence in its decisions. Intuitively a rejection rule is based on the amount of confidence a classifier has on a

given classification. The novelty rejection scheme used in this study was based on an ensemble of neural network classifiers employing a consensus scheme. The pattern is classified in the category with the majority of votes. If the number of votes the winning class receives is *g*, then a threshold is placed on *g/N* where *N* is the number of classifiers. If *g/N* is below the threshold then the pattern is rejected.

In most recognition tasks, the underlying probability distributions of the patterns are not completely known. Hence, the optimal threshold as given by Chow is no longer very useful. Fumera *et al.* (2000) tackle this problem. Their method provides an alterative, which works well even if the *a posteriori* probabilities are affected by errors. The authors suggest using multiple thresholds, one for each class. The threshold is placed on the maximum *a posteriori* probability just as in Chow(1970) but in this case each class has a different threshold. Their results using nearest neighbour and neural network classifiers show that the scheme outperforms the one based on parametric assumption.

Another improvement to Chow's original study was made by Foggia *et al.* (1999). In pattern recognition problems the idea of combining various experts with the aim of compensating the weakness of each single expert, has been widely investigated. The main problem is that the combination rule should be able to solve the conflicts i.e. when the experts disagree. In these cases the experts' final decision may be unreliable and it might be desirable to reject such patterns for more sophisticated processing. Foggia *et al.* extend the work of Cordella *et. al.*, (1995) and De Stefano *et. al.*, (2000) applying their technique to multi expert systems. The technique is also similar to that of Hansen (1997). In this paper, a reject option for the MES architecture is defined, which drops the assumption of knowing the exact values of *a posteriori* probabilities. The reject option is defined for an MES architecture with the Bayesian Combining (BC) rule. The BC rule estimates the *a posteriori* probability that the input sample belongs to each class and selects the class with the highest probability. A value between 0 and 1 can be calculated for these two reasons with values near 1 characterizing reliable classification and values near 0 indicating unreliable classifications. These two parameters can be combined in several ways and then thresholded for rejection. The threshold is computed by maximizing a function that measures the MES classification effectiveness in the considered application domain.

Some approaches have suggested the use of artificially generated anomalies that can improve the novelty detection performance of a system. Wei *et al.* (2001) suggest how artificial anomalies (novelties) can be injected into the training data to help the learner discover a boundary around the original data. To generate artificial anomalies close to the known data, a useful heuristic is to randomly change the value of one feature of an existing example while leaving the other features unaltered. Sparse regions are characterized by infrequent values of individual features. To amplify sparse regions, they proportionally create more artificial anomalies around them. The technique only works if the novel class (anomalies) do not overlap with the known classes. After the random anomalies are generated, a classification takes place and random anomalies that are misclassified as known are removed from the data. The process is repeated until a satisfactory size of stable random anomalies data is created. This system was tested on a network intrusion detection task. It was found that better performance is obtained if the training data is augmented with data of the anomalous classes. It was also found that the number of random anomalies generated is not critical for the system performance.

We now discuss some of the advanced methods of parametric statistical novelty detection including probabilistic/GMM approaches (section 2.1.1), Hidden Markov Models (section 2.1.2) and hypothesis testing (section 2.1.3). These approaches are more complex than the use of linear schemes of novelty detection (e.g. Elad *et al.*, 2002) and they are primarily aimed at data density estimation using robust statistics.

2.1.1 *Probabilistic/GMM Approaches*

Gaussian Mixture Modelling (GMM) models general distributions estimating the density using fewer kernels than the number of patterns in the training set (Odin and Addison, 2000). The parameters of the model are chosen by maximizing the log likelihood of the training data with respect to the model. This is done using optimisation algorithms such as Conjugate Gradients or re-estimation techniques such as the EM algorithm. However, GMM suffer from the curse of dimensionality in the sense that if the dimensionality of the data is high, a very large number of samples are needed to train the model. A number of studies have used GMM for novelty detection as described below.

Roberts and Tarassenko (1994) developed a robust method for novelty detection, which aims to minimize the number of heuristically chosen thresholds in the novelty decision process. The novelty detection method used is similar to those of Barnett and Lewis (1994), Bishop (1994), Tarassenko (1995), Parra *et. al.* (1995), Tax and Duin (1998), Desforges *et. al.* (1998), Brotherton *et. al.* (1998, 2000), Tarassenko *et. al.* (1999), Yeung and Chow (2002), and others, based on the density function of the training data estimated with a GMM. The parameters of the GMM are estimated with the EM algorithm. The major contribution of this paper is that the algorithm decides to add a Gaussian unit based on some automatic criterion that determines the number of Gaussians. The growth decision is based on monitoring the smallest Mahalanobis distance between a training vector and each Gaussian within the network. The growth threshold is initially set to 0 and it is progressively increased. The algorithm ensures that every Gaussian has seen each sample in the training data at least once. The maximum growth threshold found by the algorithm is also used as the novelty threshold. If the maximum posterior probability of a test vector is below this threshold then it is rejected as a novelty. The technique was tested on a medical signal-processing task to detect epileptic seizures. A total of 195 Gaussian kernels were grown by the system exhibiting very high performance rates.

Tarassenko (1995) applies his novelty detection technique to the detection of masses in mammograms. Mammography is a well-suited problem for novelty detection as often the question is whether a mass exists in the mammogram or not, and examples of abnormal tissue are often scarce compared to examples of normal tissue. The idea is to build a model of normality of the training data using only normal examples and then compare the test patterns against this model. An assumption is made that the abnormalities are uniformly distributed outside the boundaries of normality. The description of normality is made using the unconditional probability density estimation of the training data. If a test vector falls in a region of input space with a density under a pre-determined threshold then the test vector is considered to be novel. This technique is very similar to Bishop (1994), except that it tackles regions in the training space with low density but with legitimate training objects. If such regions exist, setting a global novelty threshold will fail to reject all novel cases but will reject a lot of normal cases. The solution presented here is the implementation of a local novelty threshold that depends on the density of the data in that region of input space. The space is partitioned using the *k*-means algorithm to several parts according to their input space and the density function is calculated independently within each partition. The precise number of partitions is not critical. The authors considered two ways of density estimation, Gaussian mixture models and Parzen windows and found that Parzen windows work much better due to the unavailability of a large number of training samples. The technique was tested on 120 images from the MIAS database. In all of the 40 cases the mass-like structures were correctly identified as novel except in two cases. Unfortunately a large number of false positives were also discovered.

A similar approach is taken by Tarssenko *et al.* (1999) for jet engine fault detection. The input data is initially pre-processed and a simpler model of the distribution of the input space is chosen in a

transformed space. The transform is such that the Euclidean distance in the transformed space is equal to the Mahalanobis distance in the original space. The transformation is done using two different methods, component-wise normalisation and the whitening transform. The distribution of normal vectors in the transformed space is approximated by a small number of spherical clusters selected using the *k*-means clustering algorithm ($k<5$). Each cluster radius is calculated as the average distance between the feature vectors belonging to the cluster and the cluster centre. The novelty test is based on thresholding the shortest normalised distance of the test vector to a cluster centre. The distance to the nearest cluster mean is normalised by the radius of the cluster in order to account for varying data densities in different regions of the input space. If the test vector is sufficiently far from all cluster centres then it is in a region of space with very few known training samples and hence it is deemed to be novel. The novelty threshold is set so as to accept all training samples. This method was tested on jet engine fault detection and compared with Parzen windows method used in previous study (Tarassenko *et. al.*, 1995). The method described slightly outperforms Parzen windows for the component-wise normalisation transformation.

Parra *et al.* (1995) present a new scheme of density estimation. Novelty detection is performed by finding the underlying density of the training data using this novel technique. A hyper-sphere is drawn to separate known regions from unknown regions. Novel objects should ideally fall outside this hyper-sphere. An appropriate threshold separates known from new test objects. The strength of this study is its density estimation technique that can be used with non-linear distributions or distributions for which no *a priori* information is available. The factorization of a joint probability distribution was formulated as a minimal mutual information criterion under the constraint of volume conservation. Volume conservation is implemented by symplectic maps. A Gaussian upper bound leads to a computationally efficient optimisation technique, which in turn facilitates density estimation. The method was tested for motor fault detection. The data was very high dimensional and a combination of linear PCA and symplectic maps was used to reduce the dimensionality. The novelty detection method then gave as good or slightly better results when compared to various other methods including MLP, RBF, nearest neighbour and others.

Nairac *et al.* (1997,1999) present a method for novelty detection based on a probabilistic framework applied to those tasks when only a limited amount of training data is available. It is important that the dimensionality of the data is kept relatively small because in order for a good probability model to be estimated, the number of inputs must be significantly larger than the number of dimensions. Because of the small number of samples in the training set, a GMM cannot be used. Instead, a normalising transformation is applied to the whole of the training data and a small number of spherical basis functions are then fitted to the transformed data. The authors have used two normalising methods; component-wise normalisation and a whitening transformation. The placement of the basis functions is done by a *k*-means algorithm with Euclidean distance in the transformed space. For novelty detection, the average distance between the data vectors and their corresponding basis function in the training data is computed. The novelty of a test vector is assessed by computing the shortest normalised distance to a kernel centre. This is a locally defined measure of novelty, since it is based on the distance of the nearest kernel and it is normalised by the average distance of the data within that kernel. A validation set is used to set the novelty threshold which is set as the most 'novel' point in the training set. The method was tested on a fault detection task in a jet engine. The component-wise normalisation gives much better results in recognising previously unseen normal engine data although the whitening transform is slightly better in recognising abnormal data. The authors also suggest using a Monte Carlo simulation to generate synthetic data to facilitate estimating the data density of the training data when not enough samples are present.

Tax and Duin (1998) describe three methods of rejecting outliers based on the data density distribution. Data objects in low probability areas are rejected using these methods (Barnett and

Louis, 1978; Bishop, 1994; Tarassenko *et. al.*, 1995). The techniques are based on different methods of computing data density: mixtures of Gaussians, Parzen windows and a nearest neighbour-based estimator. These methods for novelty detection are then compared with a method based on classifier instability introduced in this study. Using the mixtures of Gaussians method and Parzen windows, the density distribution of the training data is found. When the difference between the new object and the mean of the training objects is larger than three standard deviations in the training distribution, the new object is rejected. With the nearest neighbour method, a slightly different approach is followed to find the probability density. The distance of the new object and its nearest neighbour in the training set is found and the distance of this nearest neighbour and its nearest neighbour in the training set is also found. The quotient between the first and the second distance is taken as indication of the novelty of the object. The new method of novelty detection introduced here is based on classifier instability. They use a linear classifier based on maximizing Fisher's criterion and extending the two-class classifier for multiple classes. They take bootstrap samples of the same size as the original training set and train several classifiers. The outputs of these classifiers differ. The variation in the outputs indicates the diversity between different training sets. A large variation indicates that the object is hard to classify. This variation is then thresholded. The results show that this new method only outperforms the distribution probability methods when only a small training set is available. In all other cases, and especially when there is an abundance of training data, Parzen windows method is the best. However, the drawback of using Parzen windows is that a large number of samples are required for a proper estimation and the width parameter σ needs to be user-set. For the Gaussian case, the problem is selecting the correct number of kernels required which is often intuitive. The system was tested on hand written digits with good success.

Roberts (1999, 2002) proposes Extreme Value Theory (EVT) for novelty detection that concerns abnormally low or high values in the tails of data distributions. As more data is observed, the value of this extreme data changes. Knowledge of such statistics is useful in such tasks as novelty detection, outlier removal or for rejecting classification or regression of patterns that lie away from the expected statistics of some training data set. A lot of work has been done previously in this area for datasets that are known to be pure i.e. datasets that contain patterns from only one class, e.g. (Bishop, 1994; Tarassenko, 1995; Parra *et. al.*, 1995; Tarassenko *et. al.*, 1999; Brotherton *et. al.*, 1998). These methods can be used in cases where data for the normal case is abundant and easily obtainable whereas examples for the abnormal case are rare and very expensive to obtain such as in medical and fault detection domains. However, these methods become very sensitive in those cases where the normal case contains very few examples of the abnormal class. These examples are not enough for classification but enough to influence the density distribution estimation of the normal class by fitting abnormal data thus making the detection of abnormal cases very problematic. EVT forms representations for the tails of distributions. Fisher and Tippett (1928) showed that if the distribution function over a data point is to be stable as the number of samples tends to infinity then it must weakly converge under a positive affine transform. The second key theorem of Fisher and Tippett states that if the distribution is a non-degenerate limit distribution for normalised maxima then it can take only one of three forms. The first of these forms is referred to as Gumbel distribution, and used by Roberts (1999, 2002). His research is ultimately concerned with samples drawn from a distribution whose maxima distribution converges to the Gumbel form. This distribution gives a probability of observing some extreme value and the parameters can be determined using a Monte-Carlo simulation. A Gaussian Mixture Model (GMM) is used to estimate the distribution of the data, hence the EVT probability can be used directly as the EVT distribution is derived from the same range of data that the GMM is fitted on. However, only the component closest to the data point concerned (in the Mahalanobis sense) is used to calculate the EVT probability as this dominates the EVT statistics. The EM algorithm is used to estimate the parameters of the GMM using the minimum length coding to penalize models with higher complexity. The technique was tested on a tremor and an epilepsy dataset with the objective of

discriminating between normal and abnormal behaviour as well as on a noise removal task in an image (salt and pepper noise). The technique exhibited very good performance on all three tasks without the need of setting novelty thresholds that plague other techniques based on statistical novelty detection.

Yamanishi *et al.* (2000) present SmartSifter (SS), an outlier detection system from the viewpoint of statistical learning theory, which works in data mining to detect fraud, network intrusion, network monitoring etc.. The work is focused on outlier detection based on unsupervised learning of the information source. Every time a datum is input, it is required to evaluate how large the datum has deviated compared to a normal pattern. SS uses a probabilistic model as representation of an underlying mechanism for data generation. The probability density over the domain of categorical variables is found using a histogram and a finite mixture model is employed for each histogram cell to represent the probability density over the domain of continues variables. Every time a datum is input, an on-line learning algorithm is employed to update the model. The authors have developed the Sequentially Discounting Laplace Estimation for learning the histogram density over the categorical domain and the Sequentially Discounting Expectation and Maximising for learning the finite mixture for the continues domain. SS gives a score to each datum on the basis of the learned model indicating how much the model has changed after learning. A high score means that the datum is an outlier. According to the authors, the novel features of SS include the fact that SS is adaptive to non-stationary data. This is useful in the cases when drifting sources of time-series data are tackled. Added to that, a score calculated by SS has a clear statistical and information theoretic meaning. While in other works heuristics such as cost or distances such as Mahalanobis etc. are used to describe outliers, SS defines a score for a datum based on how much the model has shifted after learning it. Finally, SS is computationally inexpensive and it can deal with both categorical and continues variables. The system was successfully tested on the network intrusion database, KDD Cup, 1999.

Spence *et al.* (2001) independently developed a class of models for probability distributions of images called Hierarchical Image Probability (HIP) models. HIP constructs a tree-structured graph of the dependencies between hidden variables at different scales, and uses mixtures of multivariate Gaussians to model the local distributions of vectors of features. The hidden variables are similar to Markov Random Fields (MRF). The method is applied to mammographic images to reject those that contain cancerous regions given a set of normal images as training data. Due to the tree structure, the belief network for the hidden variables is relatively straightforward to train with an EM algorithm. The expectation step can be performed directly. The expectation is weighted by the probability of a label or a parent-child pair of labels given the image. Once the expectations are computed, the normal distribution makes the *M* step tractable. Detecting novel examples can be useful in a CAD system for generating confidence measures on the CAD output and identifying data that could be useful for future training of the model. The HIP model's generative structure enables novel examples to be identified by thresholding the log-likelihood of the models. The system was tested on mammography images but no extensive results are shown nor a comparison is made with competing systems. In addition, the authors do not explain how the novelty threshold is selected or how critical it is to the system performance.

Almost all statistical approaches dealing with novelty detection are based on modelling the density of the training data and rejecting test patterns that fall in regions of low density. However, in order for such techniques to work the training data itself needs to be either free of outliers or the outliers to be known. The most common data descriptor is the Gaussian Mixture Model (GMM). Lauer (2001) detail a method that works when the training data is corrupted with an unknown number of outliers. This approach is more robust against outliers so it can tolerate a small number of them in the training data. To calibrate the algorithm, classified validation patterns are needed or a rough estimate of the proportion of outliers in the training data. They start with the presumption that the

training data contains a small number of outliers. They model the pattern distribution as the composition of a big percentage of normal patterns and a small proportion of corrupted patterns. If λ is the proportion of anomalous patterns, P_N is the distribution of normal patterns and P_O is the distribution of anomalous patterns then the distribution of the whole training set is given as: $P(x) = (1 - l)P_N(x) + lP_O(x)$. The value λ can be interpreted as the prior probability for outlying patterns and the modelling is called the "mixture alternative". The learning task can now be decomposed into three steps: (a) Estimate $P(x)$ from the given training data; (b) Decompose $P(x)$ into $P_N(x)$ and $P_O(x)$, and (c) Decide whether x is an outlier given the probabilities $P_N(x)$, $P_O(x)$ and the prior λ. The second step is very difficult since there is no knowledge of anomalous patterns or their number. In addition, the distribution $P_O(x)$ cannot be estimated reliably due to the small number of outliers in the training data. Therefore, the second step cannot be performed directly, and instead, an assumption on the distribution $P_O(x)$ is made and the proportion λ, and $P_N(x)$ is estimated on the basis of that. The estimation of $P_N(x)$ is equivalent to the determination of the distribution's parameters. The usage of GMM is considered here and the EM algorithm is employed to estimate the parameters. The approach is run with different values of λ. The parameter λ can be optimised to resample the expected proportion of outliers. In this respect, λ is interpreted as a parameter of the algorithm that controls the sensitivity against anomalous patterns. The system was tested on artificial datasets of various complexities and a real database of medical data with patients carrying rare diseases. The results were compared with the traditional approach to novelty detection using data densities but without providing for outliers in the training set. This approach outperforms the traditional approach.

Hickinbotham and Austin (2000) describe a novelty detection method based on Gaussian Mixture Models (GMM) for sensor fault detection. The structural health of airframes is often monitored by analysing the frequency of occurrence matrix (FOOM) produced after each flight. Unfortunately, these FOOMs also get corrupted through time and corrupted FOOMs need to be filtered out. There are two classes of sensor fault. The first is a random addition of counts and the second is a shift in the response of a sensor to the load it monitors. Noise distortion is independent of the nature of the flight. They search for noise using the mean and variance of each cell in the FOOMs. By setting a threshold it is possible to count the number of cells whose value was "unlikely" using the Gaussian measure of probability. A suitable threshold can be set experimentally. For the second problem it is necessary to reduce the dimensionality of the input data. The authors used the eigenface algorithm to achieve this reduction. After the dimensionality of the training data is reduced, the authors model the distribution of FOOM data using a Gaussian basis function neural network. The EM algorithm is used to estimate the parameters of the GMM. The GMM provides a probability density function which forms the basis of the novelty measure *N*. It is useful to scale *N* so that its value lies between 0 and 1. A threshold is placed on *N* reject patterns with low probability belonging to the mixture model. The experimental results showed good performance rejecting faulty FOOMs but a large number of healthy FOOMs were also rejected. The authors state that this is probably due to the high variability of strain events between flights. It is therefore anticipated that the addition of more data will correct this problem.

Novelty detection in textual data based applications has also generated considerable interested in the past (Baker, 1999). Topic Detection and Tracking (TDT), or novelty detection, is a variant of the traditional classification problem to allow the classification of new classes. In TDT new topics emerge often which means that the classifier has to expand its setoff classes continuously. The authors cluster the unlabelled documents using a multinomial naï ve Bayes classifier estimating the parameters using the EM algorithm. The estimated probabilities are used to classify the unlabelled documents to one of the known classes or a novel class. In a novelty detection task, the parameter estimation of newly identified classes is unreliable due to a complete lack of data. The authors propose an agglomeration of the classes into hierarchy, with each node formed recursively by

pooling the data by its children. Then they use shrinkage to improve these estimates. Shrinkage linearly interpolates the parameter estimates of each leaf in the hierarchy with the parameter estimates of the leaf's ancestors. As an alternative to the EM algorithm, the authors use Deterministic Annealing (DA) for clustering. DA is based on probabilistic and information-theoretic principles that can be used to build a hierarchy and avoids many poor locally optimal solutions that can trap the EM algorithm. By constraining the entropy, they implicitly smooth the surface on which EM is hill climbing, thus reducing the number of local optima in which EM might get trapped. To build the hierarchy, the algorithm starts with infinite temperature so that no matter how many classes EM might assume, all parameter estimates converge to the same value. Thus, there is effectively only one class and it becomes the root of the hierarchy. As the temperature is lowered, the system undergoes phase transitions at which the effective number of clusters grows by splitting one node in the hierarchy into two children. The approach is extended to on-line novelty detection by assuming a hierarchy obtained either by existing data or *a priori* knowledge and determining for each subsequent document whether it is the first to report on a given topic or not. If the document is most likely in one of the new nodes then the document is labelled as new. Similarly, Hansen et al. (2000) discuss the role of Gaussian Mixture Modelling for textual data and novelty detection. The main purpose of this study is to demonstrate how well known techniques for signal/data analysis can also be used for textual information.

2.1.2 *Hidden Markov Models (HMM)*

HMMs are stochastic models for sequential data (Duda *et al.*, 2001). A given HMM contains a finite number of unobservable (hidden) states. State transitions are governed by a stochastic process to form Markov chains. At each state, some state-dependent events can be observed. The emission probabilities of these observable events are determined by a probability distribution, one for each state. To estimate the parameters of an HMM for modelling normal system behaviour, sequences of normal events collected from normal system operation are used as training data. An expectation-maximization (EM) algorithm is used to estimate the parameters. Once an HMM has been trained, when confronted with test data, probability measures can be thresholded for novelty detection.

Yeung *et al.* (2002) describe the use of HMMs for novelty detection for the application of intrusion detection based on profiling system call sequences and shell command sequences (computer security). They present two main approaches. The first one is based on Hidden Markov Models (HMM). Given a trained HMM, the sample likelihood of an observed sequence with respect to the model can be computed with the forward or backward algorithm. A threshold on the probability can discriminate between normal and abnormal behaviour. The second technique used is even simpler. The probability distribution of normal system behaviour observed over some time is modelled. A simple occurrence frequency distribution is used for this purpose. The behaviour of the system being monitored is modelled in the same way. An information-theoretic measure known as cross-entropy is used to measure how dissimilar the two distributions are. A threshold can determine whether the observed behaviour is abnormal. A validation set is used to determine the threshold. In the case of HMM, the threshold was chosen to be the minimum likelihood among all training sequences. For the second model, a cross-entropy value was computed between the entire training set and each trace in the training set. The threshold was chosen to be the maximum cross-entropy. The information-theoretic technique outperformed the HMM approach in all experiments. However, the HMM model is better suited for intrusion detection based on system calls.

2.1.3 *Hypothesis testing*

A simple statistical technique for novelty detection can be based on determining whether the test sample(s) comes from the same distribution as training data or not. Ruotolo and Surace (1997) use this approach using a *t*-test to find damaged beams. The approach described here starts by taking n_1 measurements of the first n_f natural vibration frequencies of the structure at the virgin stage i.e. as soon as the structure is built. After that, they periodically take n_2 measurements of the same natural

frequencies and use a *t*-test to compare n_1 and n_2. If the test shows significant difference between the two sets of measurements then damage is present. The approach was tested on a case study involving a continuous beam. The experimental data of the undamaged and damaged beam was obtained running a finite element program where the cracks are simulated. The test was performed at 0.05 and 0.01 significance levels and showed promising results.

2.2 *Non-parametric approaches*

Non-parametric approaches do not make any assumption on the statistical properties of data. Here we consider three approaches namely, nearest neighbour based density estimation, Parzen density estimation and string matching approaches.

2.2.1 *k*NN *based approaches*

The *k*-nearest neighbour algorithm is another technique for estimating the density function of data (Odin and Addison, 2000). This technique overcomes some of the problems of Parzen window in that it does not require a smoothing parameter. Instead, the width parameter is set as a result of the position of the data point in relation to other data points by considering the *k*-nearest patterns in the training data to the test pattern. The problem with this technique is that for large sized datasets, a large number of computations have to be performed. For novelty detection the distribution of normal vectors is described by a small number of spherical clusters placed by the *k*-nearest neighbour technique. Novelty is assessed by measuring the normalised distance of a test sample from the cluster centres. A number of studies have used such an approach to novelty detection as detailed below.

Hellman (1970) used the Nearest Neighbour (NN) classifier for rejecting patterns with higher risk of being misclassified. This was in the same spirit as Chow (1970) who used a threshold on the *a posteriori* probabilities of a Bayes optimum recognition system rejecting patterns with low probability. Chow however assumed that for a given data, *a priori* probabilities and conditional probability density are known. In most practical applications however these statistics are unknown and have to be inferred from a set of labelled examples. The advantage of this approach is that it makes no assumptions concerning these underlying statistics. Moreover, Cover and Hart (1967) have shown that as the number of training samples tends to infinity the nearest neighbour risk is no greater than twice the Bayes risk, regardless of the underlying statistics. Hellman explained that using the single NN there is not enough information to reject on some samples and not on others. More information is necessary and it is given by considering two NNs. If both come from the same class then the pattern is classified otherwise it is rejected. Hellman showed that going from the single NN to the new rule, the increase in reject rate is twice the decrease in probability of error. Thus, if errors are at least twice as costly as rejects, the new rule always has lower total cost than the single NN rule. This rule can be extended to examine *k* NNs and classify only when all the neighbours agree. This approach is very conservative and hence the author proposed an alternative: examine the *k* NNs and if all neighbours agree classify the test pattern otherwise reject it. It has to be said here that the author does not claim this to be a novelty detection approach, but merely a method for rejecting patterns with high misclassification risk. It can however work for novelty detection if the novel classes induce high confusion in the classifier; otherwise the novel classes will be erroneously classified to one of the training classes.

Guttormsson *et al.* (1999) present a novelty detection method applied to rotor fault detection. The training data is first subjected to simple outlier removal to ensure pureness. Any pattern that lies more than three standard deviations from the average is removed from the set. For novelty detection, a surface is imposed around the set of healthy signals. If a new signal falls outside this surface it is deemed to be novel. Surfaces that can be placed around the healthy points include a spherical boundary, an elliptical boundary, a rectangular boundary formed by the extrema of the data, or min-max surface and nearest neighbour boundaries. In the first two methods, a new pattern

is compared to the centre of either the hyper-sphere using the Euclidean distance or the ellipse using the Mahalanobis distance. For the min-max technique, the smallest possible box containing all the healthy data is used. The dimensions of the box are determined by the minima and maxima of the signature signals. The nearest neighbour novelty detector allows for more general data topology. Here, minimum Euclidean distances are found between each point and its closest neighbour. The distance proportional to the maximum of these distances is then used as a decision parameter. Every incoming point is compared to every point in the healthy set. If the new point is at a greater distance from each of the healthy points than the decision parameter then a novelty is declared. The radius of the hyper-sphere and the ellipse controls the trade-off between novelty detection and false alarms. Similarly the ball drawn around each training point in the nearest neighbour method and the min-max parameters in the hyper-box methods increase the method's sensitivity to novelty. The experimental results showed that the elliptical surface is the most suitable for novelty detection.

A number of techniques model the density of the training data and use it for novelty detection, e.g. (Barnett and Louis, 1978; Bishop, 1994; Tarassenko, 1995; Parra *et. al.*, 1995; Tax and Duin, 1998; Desforges *et. al.*, 1998; Brotherton *et. al.*, 1998; Tarassenko *et. al.*, 1999; Yeung and Chow, 2002). This requires a large number of samples to overcome the curse of dimensionality. In fact most other novelty detection techniques require a large amount of data to form the 'normal' class especially in high dimensions. The simplest model when just a little amount of data is available is the unimodal normal distribution. Tax and Duin (2000) suggest modelling the probability density of the data with a unimodal normal distribution and threshold the probability density accepting 95% of the data placing a threshold on the Mahalanobis distance. Instead of modelling the complete probability density, an indication can be obtained by comparing distances. This method is based on the local density of the test object and the nearest neighbour in the training set. The distance between the test pattern and its nearest neighbour in the training data is found along with the distance of the neighbour and it's own nearest neighbour. The quotient between the two distances is an indication of novelty. The Euclidian distance is used for this purpose. This method is very useful for distributions with relatively fast decaying probabilities. The techniques were tested on both real and artificial data and found to be very useful when very little amount of training data exist (less than 5 samples per feature).

Jiang *et al.* (2001) propose a two-phase clustering algorithm for outlier detection based on a modified *k*-means algorithm and a Minimum Spanning Tree (MST). The *k*-means algorithm is modified as to calculate the minimum distance between any pair of cluster centres. If the distance of a pattern and its closest cluster is larger than this distance, then the pattern is assigned to a new cluster. In the extreme case, each pattern will form its own cluster and therefore so an upper limit *k*max is defined. When *k*max is reached, two nearest clusters are merged. The cluster centres defined with the modified *k*-means algorithm are regarded as nodes, which are used to form an MST based upon the distance between every two nodes. After the MST is constructed, the longest edge of a tree is removed from the forest and it is replaced with two newly obtained sub trees. The small clusters (the tree with less number of nodes) are selected and regarded as outliers. Three different datasets were used to compare this technique with the traditional *k*-means clustering algorithm. In all of their experiments, the new technique outperformed baseline techniques tested. The datasets used include the iris data, sugar cane breeding data and e-mail log data.

Yang and Liu (1999) describe two approaches to novelty detection in document classification: Linear Least Squares Fit (LLSF) and *k*-Nearest Neighbour (*k*-NN) classifier approach. The *k*-NN algorithm is very simple. The *k* nearest neighbours of a test pattern are found in the training set. The categories of the *k* nearest neighbours are used to weight the category candidates. The similarity score of each neighbour is used as a weight. If several neighbours share the same category, then the per-neighbour weights of that category are added together, and the resulting weighted sum is used as the likelihood score of that category with respect to the test pattern. A threshold set using cross-

validation can be applied to determine whether a test pattern is sufficiently 'novel' to be rejected. In LLSF, a multivariate regression model is automatically learnt from the training data and its categories. By solving a linear least-squares fit on the training pairs (input-output) of vectors, one can obtain a matrix of regression coefficients. The solution matrix defines a mapping from an arbitrary pattern to a vector of weighted categories. By sorting these category weights, a ranked list of categories is obtained for the pattern. By thresholding on these category weights, category assignments can be obtained. Each category has a specific threshold determined using cross-validation. A minimum on each threshold can be imposed for novelty detection and pattern rejection.

Yang *et al.* (2002) describe a very simple novelty detection method applied to document classification. Training data from old events is used to learn useful statistics for the prediction of new (novel) events. Their approach consists of the following steps: classifying documents into broad topics each of which consists of multiple events, identifying named entities, optimising their weight relative to normal words for each topic, and computing a stop-word list per topic. Finally, they measure the novelty of a new document conditioned on the system-predicted topic for that document. The algorithm is very simple. When a new document arrives, it is compared with all the documents available. If its nearest neighbour in its past has a cosine similarity score below a threshold, then the document is labelled as novel, meaning that it is the first story of a novel event otherwise it is labelled as old and added to the history. The threshold is set using cross-validation. The algorithm works at two-levels. At the first level, a classifier determines the broad topic of the arrived document and at the second level the novelty detector decides if the document describes a new event or an old event. The method was applied to a benchmark document database and showed very good performance. The technique is very simplistic and will probably fail in most other domains that exhibit high complexity between the various objects. It is however an interesting approach to novelty detection.

2.2.2 *Parzen density estimation*

Parzen windows method (Duda *et al.*, 2001) can be used for non-parametric data density estimation. Yeung and Chow (2002) follow a well-established novelty detection approach, based on estimating the density of the training data and rejecting patterns (similar to Bishop, 1994; Tarassenko, 1995; Parra *et. al.*, 1995; Tax and Duin, 1998; Desforges *et. al.*, 1998; Brotherton *et. al.*, 1998; Tarassenko *et. al.*, 1999). The authors apply their technique on an intrusion detection problem. The authors have chosen Gaussian kernel functions for two reasons. First, the Gaussian function is smooth and hence the density estimation also varies smoothly and second, if a radially symmetrical Gaussian is assumed, the function can be completely specified by a variance parameter only. The novelty threshold is set using a separate training set called 'threshold determination set' and it is applied on the unconditional probability $p(x)$ of a test pattern x based on the modelled distribution. The technique was tested using the dataset from KDD Cup, 1999.

2.2.3 *String Matching approaches*

String matching approaches are based on treating training data as templates represented by a string (vector of features) and then computing some measure of dissimilarity between training and test data. Forrest *et al.* (1994) present a method for solving the problem of distinguishing self from non-self using a change-detection algorithm based on the way the natural immune system achieves the same task. The self-data is converted to binary format forming a collection S. Then a large number of random strings are generated forming a set R_0. Strings from R_0 are matched against the strings in S and those that match are eliminated. Since perfect matching is extremely rare, the matching criterion is relaxed so as to consider only r contiguous matches in the strings. Once R_0 is created, new patterns are converted to binary and matched against R_0. If a match is found, then the new pattern belongs to non-self and is rejected. A number of experiments were performed designed to

test various aspects of the detection system. The system was tested on network intrusion tasks and found to perform extremely well.

A further extension to this work is provided by Dasgupta and Forrest (1996) and Dasgupta and Nino (2000) whose work is biologically inspired (based on immuno-computing concepts, Tarakov and Skormin, 2002). In our bodies novelty detection is carried out by T-cells that have receptors in their surface that can detect foreign proteins. The body releases a large number of T-cells but only allows those that do not match any of the body's own cells to circulate. If a T-cell matches any cell in the body then that cell is deemed as foreign and it is dealt with. The novelty detection technique presented here works in a similar way. A sufficient subset of the training data is taken and the analogue values are made discrete by sampling. Each point is assigned an integer value and then converted to binary form. Thereafter, a large set of strings, called detectors, is generated that do not match the strings obtained by the training data. If a new data point, after being encoded, matches any of the detectors then a deviation from the normal system behaviour is evaluated and it is treated as novel data. The matching is performed based on a matching threshold r that sets the number of bits that have to match before two strings are deemed similar. One major concern, as identified by the authors, of this system is the matching threshold. This has to be data specific and its correct setting is essential for satisfactory system performance. With larger r the detectors become too sensitive to any novelty in the data whereas small r might not result on a reasonable size of detector set. This technique was tested on tool breakage detection and the results on an average of 50 runs showed very good performance in detecting abnormal behaviour.

Further work in this area extends the previous two studies (Dasgupta and Gonzalez, 2001). This paper extends the idea presented in previous work of Dasgupta and Forrest (1996) and Dasgupta and Nino (2000) to multi-class approach. Specifically the non-self (unknown) space will be further classified into multiple subclasses to determine the level of abnormality. The technique is implemented on a computer intrusion task. In an anomaly detection system it is an acceptable assumption that the normal operation of the system can be characterized by a series of observations over time. Also, normal system behaviour generally exhibits stable patterns when observed over a period of time. This is the common ground in many fault detection systems including those proposed by (Tarassenko, 1995; Japkowicz, 1995; Dasgupta and Forrest, 1996, 2000; Tarassenko, 1999; and Campbell and Bennett, 2001). According to Dasgupta and Gonzalez, a naïve approach to the task is to determine the minimum and maximum values of the monitored parameters and measure the abnormality as a deviation from these values. However, such an approach will not consider the fact that normality is dependent on time and values that might be acceptable at a given time might not be acceptable at a different time. Added to that, the notion of normality depends on the correlation and interaction of various parameters (features). In this paper, a sliding window is used for pattern characterization and the normal behaviour of the system is represented by a subspace called Self and its compliment Non-Self. Two approaches are described here for fault detection, positive characterization and negative characterization. Positive characterization is a nearest neighbour technique that records the Euclidean distance between a test vector and its nearest neighbour in the Self subspace. A user-set value determines the allowable variability in the Self subspace. If the distance exceeds this value then the vector is deemed to be abnormal. The technique is implemented using KD-Tree for faster querying. The negative characterization approach is more in tune with Dasgupta and Forrest (1996) and Dasgupta and Nino (2000) but implemented using Genetic Algorithms (GAs) to build a representation of the Non-Self subspace using the Self subspace as input. GAs are used to evolve rules to cover the Non-Self subspace, although the shape of neither Self or Non-self subspaces is known *a priori* (the patterns that belong to Self can be used to approximate those subspaces). The Genetic Algorithm attempts to evolve 'good' rules that cover the Non-Self space. The goodness of rule is determined by various factors: the number of normal samples that cover the space, its area and the overlap with other rules. This is a multi-objective, multi-modal optimisation problem. The objective is to find not only a single

solution but a number of solutions that cooperatively solve the problem. Since the covering of the Non-Self space is accomplished by a set of rules, it is necessary to evolve multiple rules. A sequential niching algorithm is employed to evolve these different rules. During system operation, if a test vector falls in the Non-Self space then it is deemed to be abnormal. The two techniques were tested and compared on a computer intrusion detection system. The data was obtained from the MIT-Lincoln Lab. The attack free data was used for training and the system was tested using both systems. The negative characterization approach was clearly more efficient (in time and space) compared to the positive characterization although the positive characterization exhibited more precise results.

Finally, Dasgupta and Majumdar (2002) extend the above studies for multi-dimensional data. The technique proposed is identical to the one described in those previous papers except that the multidimensional data is first passed through PCA for dimensionality reduction discarding features that accounted for less than 20% variability selecting only two dimensions. The two dimensions were binary encoded as before but unlike those studied the binary strings were also Gray Coded. The system was tested on a network anomaly detection task and although good overall results were obtained, some anomalies went undetected.

2.2.4 *Clustering approaches*

Clustering based approaches are aimed at partitioning data into a number of clusters, where each data point can be assigned a degree of membership to each of the clusters. If the degree of membership is thresholded to suggest if a data point belongs or not to a cluster, novelty can be detected when a sample belongs to none of the available classes.

Pizzi *et al.* (2001) describe EvIdent, a data analysis software for quickly detecting, investigating and visualizing novel events in a set of images as they evolve in time and/or frequency. This work follows the earlier research by Scarth *et al.* (1995). For instance, in magnetic resonance neuro-images, novelty may manifest itself as neural activations in a time course. Conventional data analysis methods of fMRIs assume that a model of the requisite function is available (for example, a brain's response to a designed cognitive or motor stimulus) and that the validity of this model can be tested by statistical methods of inference. However, in the case of neuroscience, researchers are probing brain function with increasingly complex cognitive experiments. This complexity demands that any model validation approach should be versatile, adaptive, data-driven and model-free. Therefore, novelty detection is a prime candidate in solving this problem. EvIdent clusters time courses within the volumetric data, using a much-enhanced variant of Bezdek's fuzzy *C*-means algorithm (Bezdek *et al.*, 1984). The time courses are separated in such a way that the intra-cluster distance is minimized, while simultaneously maximizing the inter-cluster distances. The fuzzy index m controls the fuzziness of the cluster portions: as m approaches 1, the fuzzy *C*-means algorithm converges to a classical hard means algorithm. As m approaches ∞, all cluster centroids tend towards the centroid of all time courses. The use of the fuzzy *C*-means as opposed to the classical *k*-means algorithm is dictated by the ability of the former to avoid getting stuck to local minima of the objective function it tries to optimise. Fuzzy cluster analysis produces for each cluster a membership map for each anatomical slice of the volumetric data. The membership map is an image representing the degree of membership of each active voxel to each cluster. By thresholding a membership map, those voxels that belong to the cluster with at least the user-specified level of membership can be identified.

Yang *et al.* (1998) attempt to automatically detect novel events from a temporally ordered stream of news stories, either retrospectively or as the stories arrive. The objective is to identify stories in several continuous news streams that belong to previously unidentified events. This can be done in an on-line fashion i.e. as the events occur or an accumulated collection. In retrospective event detection, stories are grouped together where each cluster uniquely identifies an event. In on-line

event detection each document is labelled as it arrives in sequence with a *new* or *old* flag indicating whether or not the document is the first story discussing a novel event. Two clustering approaches are investigated: an agglomerative (hierarchical) algorithm based on group-average clustering (GAC), and a single pass algorithm (INCR) that generates a non-hierarchical partition of the input collection. The former is appropriate for retrospective event detection whereas the latter can be used for both. A story is represented using a vector of weighted terms. The normalised vector sum of documents in a cluster is used to represent the cluster and it is called a prototype or centroid. The standard cosine similarity measure is used to describe the similarity of a cluster centroid and a document. GAC is an agglomerative algorithm that maximizes the average similarity between document pairs in the resulting clusters. At each iteration it divides the current set of clusters into buckets and does local clustering within each bucket. The process is repeated and generates clusters at higher and higher levels until a predefined number of clusters are obtained. The input to the algorithm is a set of documents and the output is a forest of cluster trees with the number of trees specified by the user. Clusters are produced by growing a binary tree using the bottom up approach. Novelty detection is used in the case of single-pass clustering. The algorithm sequentially processes the input documents, one at a time, and grows clusters incrementally. A new document is classified to its most similar cluster if the similarity exceeds a predefined threshold otherwise it becomes the seed for a new cluster. By adjusting the threshold one can obtain clusters at different levels of granularity.

3. Conclusion

In this paper we have presented a survey of novelty detection using statistical approaches. Most of such research is driven by modelling data distributions and then estimating the probability of test data to belong to such distributions. In such model-based approaches, one does need to specify or make assumptions on the nature of training data. In addition, the amount and quality of training data becomes very important in the robust determination of training data distribution parameters. Statistical approaches however are cheap to compute and straight-forward in their explanation of the techniques used. Their main competition for the novelty detection task comes from a variety of neural networks, something we discuss in our successive paper.

References

1. L.D. Baker, T. Hofmann, A.K. McCallum and Y. Yang, "A hierarchical probabilistic model for novelty detection in text", Technical Report, 1999.
2. V. Barnett and T. Lewis, *Outliers in statistical data*, John Wiley, 1994.
3. J.C. Bezdek, R. Ehrlich, and W. Full, "FCM: the fuzzy c-means clustering algorithm", Computers and Geosciences, vol. 10, pp. 191-203, 1984.
4. C. Bishop, "Novelty detection and neural network validation", Proc. IEE Conference on Vision and Image Signal Processing, pp. 217-222, 1994.
5. T. Brotherton, T. Johnson and G. Chadderdon, "Classification and novelty detection using linear models and a class dependent- elliptical basis function neural network", Proc. IJCNN Conference, Anchorage, May, 1998.
6. C. Campbell and K.P. Bennett, "A linear programming approach to novelty detection", Advances in NIPS, vol. 14, MIT Press, Cambridge, MA, 2001.
7. G.A. Carpenter, M.A. Rubin and W.W. Streilein, "ARTMAP-FD: familiarity discrimination applied to radar target recognition", Proc. International Conference on Neural Networks, vol. III, pp. 1459-1464, 1997.
8. C.K. Chow, "On optimum recognition error and reject tradeoff", IEEE Transactions on Information Theory, vol. IT-16, no. 1, pp. 41-46, January, 1970.
9. L.P. Cordella, C. De Stefano, F. Tortorella and M. Vento, "A method for improving classification reliability of multilayer perceptrons", IEEE Transactions on Neural Networks, vol. 6, no. 5, pp. 1140-1147, 1995.

10. T. Cover and P. Hart, "Nearest neighbor pattern classification", IEEE Transactions in Information Theory, vol.13, pp. 21-27, 1967.
11. M.J. Desforges, P.J. Jacob and J.E. Cooper, "Applications of probability density estimation to the detection of abnormal conditions in engineering", Proc. Institute of Mechanical Engineers, vol. 212, pp. 687-703, 1998.
12. D. Dasgupta and F.A. Gonzalez, "An immunogenetic approach to intrusion detection", Division of Computer Science, University of Memphis, Tech. Report CS-01-001, 2001.
13. D. Dasgupta and S. Forrest, "Novelty-detection in time series data using ideas from immunology", Proc. International Conference on Intelligent Systems, Reno, Nevada, 1996.
14. D. Dasgupta and F. Nino, "A comparison of negative and positive selection algorithms in novel pattern detection", Proc. IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, pp. 125–130, 2000.
15. D. Dasgupta and N. S. Majumdar, "Anomaly detection in multidimensional data using negative selection algorithm", Proc. IEEE Conference on Evolutionary Computation, pp. 1039-1044, Hawaii, May 2002.
16. R.O. Duda, P.E. Hart and D.G. Stork, *Pattern classification*, Wiley, 2001.
17. M. Elad, Y. Hel-Or and R. Keshet, "Rejection based classifier for face detection", Pattern Recognition Letters, vol. 23, pp. 1459-1471, 2002.
18. R.A. Fisher, and L.H.C. Tippett, "Limiting forms of the frequency distribution of the largest and smallest member of a sample", Proc. Camb. Phil. Soc. 24: 180-190, 1928.
19. P. Foggia, C. Sansone, F. Tortorella and M. Vento, "Multiclassification: reject criteria for the Bayesian combiner", Pattern Recognition, vol. 32, no. 8, pp. 1435-1447, 1999.
20. S. Forrest, A. S. Perelson, L. Allen and R. Cherukuri, "Self-non-self discrimination in a computer", Proc. IEEE Symposium on Research in Security and Privacy, pp. 202-212, Oakland, CA, May 1994.
21. G. Fumera, F. Roli and G. Giacinto, "Reject option with multiple thresholds", Pattern Recognition, vol. 33, pp. 2099-2101, 2000.
22. R.S. Guh, F Zorriassatine and J.D.T. Tannock, "On-line control chart pattern detection and discrimination -a neural network approach", Artificial Intelligence in Engineering, vol. 13, pp. 413-425, 1999.
23. S. E. Guttormsson, R. J. Marks II, M. A. El-Sharkawi, "Elliptical novelty grouping for on-line short-turn detection of excited running rotors", IEEE Transactions on Energy Conversion, vol. 14, No. 1, March 1999.
24. L. K. Hansen, C. Liisberg and P. Salamon, "The error-reject tradeoff", Open Systems and Information Dynamics, vol. 4, pp. 159-184, 1997.
25. L. K. Hansen, S. Sigurdsson, T. Kolenda, F. A. Nielsen, U. Kjems and J. Larsen: "Modeling text with generalizable Gaussian mixtures", Proc. of IEEE ICASSP'2000, vol. 6, pp. 3494-3497, 2000.
26. M.E. Hellman, "The nearest neighbour classification with a reject option", IEEE Transactions on Systems Science and Cybernetics, vol. 6, no. 3, pp. 179-185, July 1970.
27. S.J. Hickinbotham and J. Austin, "Neural networks for novelty detection in airframe strain data", Proc. IEEE IJCNN, 2000.
28. N. Japkowicz, C. Myers and M. Gluck, "A novelty detection approach to classification", Proc. of 14th IJCAI Conference, Montreal, pp. 518-523, 1995.
29. M. F. Jiang, S. S. Tseng, C. M. Su, "Two-phase clustering algorithm for outliers detection", Pattern Recognition Letters, vol. 22, pp. 691-700, 2001.
30. S.P. King, D.M. King, P. Anuzis, K. Astley, L. Tarassenko, P. Hayton, S. Utete, "The use of novelty detection techniques for monitoring high-integrity plant", Proc. 2002 International Conference on Control Applications, vol. 1, pp. 221–226, 2002.
31. E.M. Knorr, R.T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications", VLDB Journal, 8(3-4) pp. 237-253, 2000.

32. M. Lauer, “A mixture approach to novelty detection using training data with outliers”, in Luc De Raedt, Peter Flach (eds), Proc. 12th European Conference on Machine Learning, pp. 300-311, Springer, 2001.
33. J. Laurikkala, M. Juhola, E. Kentala, “Informal identification of outliers in medical data”, Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2000), Berlin, August 2000.
34. C. Manikopoulos and S. Papavassiliou, “Network intrusion and fault detection: a statistical anomaly approach”, accepted for publication in IEEE Communications Magazine, October 2002.
35. G. Manson, G. Pierce, K. Worden, T. Monnier, P. Guy, K. Atherton, “Long term stability of normal condition data for novelty detection”, Proc. 7th International Symposium on Smart Structures and Materials, California, 2000.
36. G. Manson, G. Pierce and K. Worden, “On the long-term stability of normal condition for damage detection in a composite panel”, Proc. 4th International Conference on Damage Assessment of Structures, Cardiff, UK, June 2001.
37. G. Manson, “Identifying damage sensitive, environment insensitive features for damage detection”, Proc. IES Conference, Swansea, UK, 2002.
38. A. Nairac, T. Corbett-Clark, R. Ripley, N. Townsend and L. Tarassenko, “Choosing an appropriate model for novelty detection”, Proc. 5th IEE International Conference on Artificial Neural Networks, Cambridge, pp. 227-232, 1997.
39. A. Nairac, N. Townsend, R. Carr, S. King, P. Cowley and L. Tarassenko, “A system for the analysis of jet engine vibration data”, Integrated Computer Aided Engineering, vol. 6, pp. 53-65, 1999.
40. T. Odin and Addison D., “Novelty detection using neural network technology”, Proc. COMADEN conference, 2000.
41. L. Parra, G. Deco and S. Miesbach, “Statistical independence and novelty detection with information preserving non-linear maps”, Neural Computation, vol. 8, no. 2, pp. 260-269, 1995.
42. N. J. Pizzi, R. A. Vivanco and R. L. Somorjai “EvIdent: a functional magnetic resonance image analysis system”, Artificial Intelligence in Medicine, vol. 21, pp. 263-269, 2001.
43. S. Roberts and L. Tarassenko, “A probabilistic resource allocating network for novelty detection”, Neural Computation, vol. 6, pp. 270-284, 1994.
44. S.J. Roberts, “Novelty detection using extreme value statistics”, IEE Proc. on Vision, Image and Signal Processing, vol. 146, issue 3, pp. 124-129, 1999.
45. S.J. Roberts, “Extreme value statistics for novelty detection in biomedical signal processing”, Proc. 1st International Conference on Advances in Medical Signal and Information Processing, pp. 166-172, 2002.
46. R. Ruotolo and C. Surace, “A statistical approach to damage detection through vibration monitoring”, Proc.5th Pan American Congress of Applied Mechanics, Puerto Rico, 1997.
47. R. Saunders and J.S. Gero, “The importance of being emergent”, Proc. Artificial Intelligence in Design, 2000.
48. G. Scarth, M. McIntyre, B. Wowk, and R. Somorjai, “Detection of novelty in functional images using fuzzy clustering”, Proc. 3rd Meeting ISMRM, Nice, France, p. 238, 1995.
49. S. Singh and M. Markou, “An approach to novelty detection applied to the classification of image regions”, IEEE Transactions on Knowledge and Data Engineering, (in press, 2003).
50. C. Spence, L. Parra and P. Sajda, “Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model”, IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, MMBIA 2001, pp. 3-10, 2001.
51. C.D. Stefano, C. Sansone and M. Vento, “To reject or not to reject: that is the question - an answer in case of neural classifiers”, IEEE Transactions on Systems, Man and Cybernetics-Part C, IEEE Comp. Press, New York, vol. 30, no. 1, pp. 84-94, 2000.

52. A.O. Tarakanov, V.A. Skormin, “Pattern recognition by immunocomputing”, Proc. of the 2002 Congress on Evolutionary Computation, CEC '02., vol. 1, pp. 938 –943, 2002.
53. L. Tarassenko, “Novelty detection for the identification of masses in mammograms”, Proc. 4th IEE International Conference on Artificial Neural Networks, vol. 4, pp. 442-447, 1995.
54. L. Tarassenko, A. Nairac, N. Townsend and P. Cowley, “Novelty detection in jet engines”, IEE Colloquium on Condition Monitoring, Imagery, External Structures and Health, pp. 41-45, 1999.
55. D. M. J. Tax and R.P.W. Duin, “Outlier detection using classifier instability”, In Advances in Pattern Recognition, the Joint IAPR International Workshops, pp. 593-601, 1998.
56. D.M.J. Tax and R.P.W. Duin, “Data description in subspaces”, in International Conference on Pattern recognition, vol. 2, Barcelona, 2000.
57. A. Webb, *Statistical pattern recognition*, Arnold, 1999.
58. F. Wei, M. Miller, S.J. Stolfo, L. Wenke, P.K. Chan, “Using artificial anomalies to detect unknown and known network intrusions”, Proc. IEEE International Conference on Data Mining, ICDM 2001, pp. 123–130, 2001.
59. K. Yamanishi, J. Takeuchi, and G. Williams. “On-Line unsupervised outlier detection using finite mixtures with discounting learning algorithms”, Proc. 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 320-324, Boston, MA, USA, August 2000.
60. Y. Yang, T. Pierce and J. Carbonell, “A study on retrospective and on-line event detection”, Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 28-36, 1998.
61. Y. Yang and X. Liu “A re-examination of text categorization methods”, Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42-49, 1999.
62. Y. Yang, J. Zhang, J. Carbonell and C. Jin, “Topic-conditioned novelty detection”, International Conference on Knowledge Discovery and Data Mining, July 2002.
63. D.Y. Yeung and Y. Ding, “Host-based intrusion detection using dynamic and static behavioral models”, Pattern Recognition, vol. 36, pp. 229-243, 2002.
64. D.Y. Yeung and C. Chow, “Parzen window network intrusion detectors”, Proc. International Conference on Pattern Recognition, 2002.