Review of Novelty Detection Methods

Dubravko Miljković
Hrvatska elektroprivreda
Vukovarska 37, 10000 Zagreb
dubravko.miljkovic@hep.hr

***Abstract* - Novelty detection is the identification of new or unknown data or signals that a machine learning system is not aware of during training. Novelty detection methods try to identify outliers that differ from the distribution of ordinary data. This paper is a short review of novelty detection and its methods.**

I. INTRODUCTON

Novelty detection refers to the identification of novel or abnormal patterns embedded in a large amount of normal data. Novelty (anomaly, outlier, exception) is a pattern in the data that does not conform to the expected behavior. The goal of novelty detection is identifying abnormal system behaviors which are not consistent with the normal state of a system [1,2,3,4,5]. Novel events occur relatively infrequently, but can have very significant consequences to overall system operation. A general design of a framework for novelty detection is presented in Fig. 1., [2]. Design combines knowledge disciplines (various mathematical and algorithmic concepts applied to novelty detection) and application domains (system expertise can be utilized e.g. for selecting features from original data).

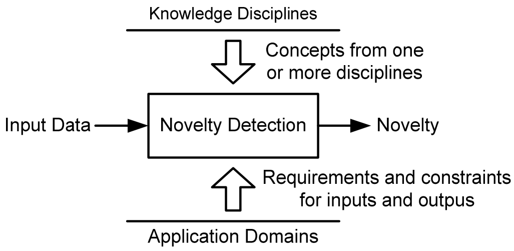
Knowledge Disciplines
Concepts from one or more disciplines
Input Data
Novelty Detection
Novelty
Requirements and constraints for inputs and outpus
Application Domains

Fig. 1. Framework of novelty detection

Input data passes through several phases: preprocessing that remove artifacts from the data, feature extraction that represent input signals using a smaller set of quantities, termed features that reduce dimensionality of input data, construction of feature vectors and normalization (component wise normalization, i.e. zero-mean, unit-variance transform). Obtained feature vectors are forwarded to novelty detection method and information about novelty is given as a final result. Normal and abnormal data in feature space are illustrated in Fig. 2. During normal operation of a system features lie within normal region. Abnormal operation of a system is detected by novel points outside normal region.

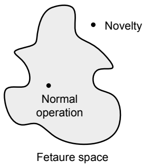

Fig. 2. Novelty within feature space

II. NOVELTY PATTERNS

Novelty patterns, based on its composition and relation to the rest of the data, can be divided in three groups [1,2].

A. Point pattern

No structure is assumed among data instances, individual instances are novel, Fig. 3.

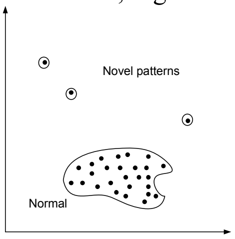

Fig. 3. Example of point novelty

B. Collective pattern

In this case novelty is a collection of related data instances (novelty is a subset of data), Fig. 4.

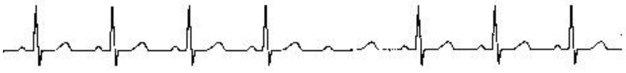

Fig. 4. Example of collective novelty

C. Contextual pattern

Novelty is an individual data instance within specific context, (conditional novelty, e.g. event at t_2 is preceded by event at t_1, within specified interval), Fig. 5.

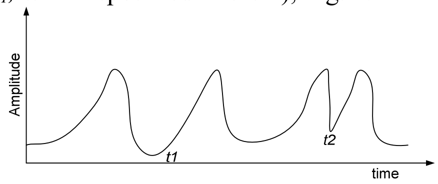

Fig. 5. Example of contextual novelty

III. NOVELTY DETECTION METHODS

Taxonomy of novelty detection is presented in Fig. 6., [1].

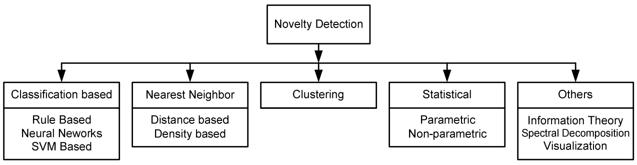

Fig. 6. Taxonomy of novelty detection methods

A. *Classification based*

In classification based novelty detection classification model is built for normal (and anomalous but rare) patterns based on labeled training data, and such model is used to classify each new unseen patterns. Common classification based approaches are listed bellow [1,2,3,4].

1) *Rule based*

Rule based techniques generate rules which either capture the normal behavior of a system. Any sequence of data deviating from these rules would be considered a novelty, [2,6], approach is illustrated in Fig. 7. from [6].

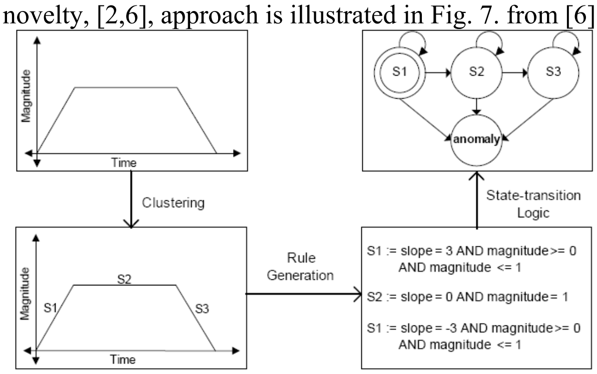
Magnitude
Time
Clustering
Magnitude
S1
S2
S3
Time
Rule Generation
S1 := slope = 3 AND magnitude >= 0 AND magnitude <= 1
S2 := slope = 0 AND magnitude = 1
S1 := slope = -3 AND magnitude >= 0 AND magnitude <= 1
State-transition Logic
S1
S2
S3
anomaly

Fig. 7. Rule based novelty detection

2) *Neural networks*

The basic idea is to train the neural network on the normal training data and then detect novelties by analyzing the response of the trained neural network to a test input. If the network accepts a test input, it is normal and if the network rejects a test input, it is novelty. Neural networks have ability to generalize (the outputs of the network approximate target values given inputs that are not in the training set), [7]. Main neural architectures employed in novelty detection are:

- Multilayer Perceptron (MLP)

A multilayer perceptron is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs, Fig. 8. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, [7].

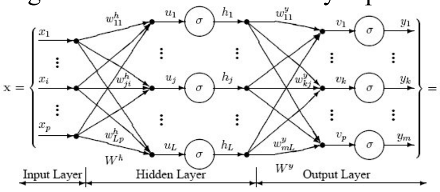

Fig. 8. Multilayer Perceptron (MLP)

The goal of the training process is to find the set of weight values that will cause the output from the neural network to match the actual target values as closely as possible. Training is accomplished by backpropagation algorithm, [7], or its version quickpropagation.

- Self-organizing map (SOM)

Self-organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically 2-D), discretized representation of the input space of the training samples, called a map. Nodes are usually arranged in a hexagonal or rectangular grid with regular spacing, Fig. 9. SOM use a neighborhood function to preserve the topological properties of the input space, [8]. After training the map contains the reference vectors of the input data. During test input data is presented to a network and there will be one winning neuron [2,4]. Departure from normality is detected by activation of neuron outside normal region of the map.

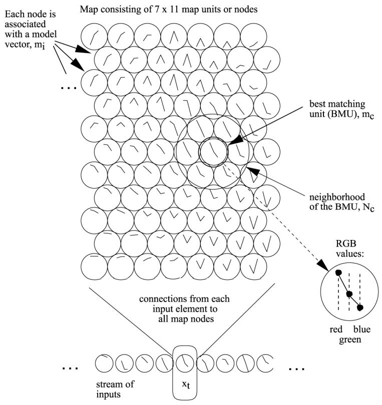
Each node is associated with a model vector, m_i
Map consisting of 7 x 11 map units or nodes
best matching unit (BMU), m_c
neighborhood of the BMU, N_c
RGB values:
red blue green
connections from each input element to all map nodes
stream of inputs
x_t

Fig. 9. Self-organizing map (SOM)

- Habituation based

Habituation is a reversible decrement in behavior response to a stimulus that is seen repeatedly without any ill effects. Habituation based approach is similar to Self-Organizing Maps. The habituation based networks tend to ignore older patterns and give more weight to newly learnt instances, by using memory leaking neurons, [2, 9]

- Neural trees

The network consists of a set of perceptrons functionally organized in a binary tree (neural tree), as illustrated in Fig. 10. A neural tree works like unsupervised hierarchical clustering algorithm, [10]. Neural tree is a hierarchical quantization of the training data into equiprobable regions.

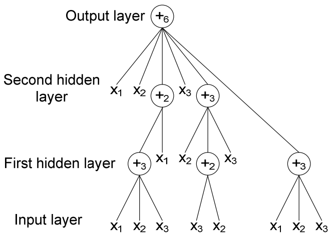
Output layer
$+_6$
Second hidden layer
x_1 x_2 $+_2$ x_3 $+_3$
First hidden layer
$+_3$ x_1 x_2 $+_2$ x_3 $+_3$
Input layer
x_1 x_2 x_3 x_3 x_2 x_1 x_2 x_3

Fig. 10. Neural tree

Neural tree performs reasonably well while being much faster than any of the other competitive learning algorithms. The tree constructed from training data servers as a reference tree. Another tree is built for the testing data. Novelty is detected when these two trees differ too much (Kullbach-Liebler Divergence or log-likelihood ratio).

- Auto-associative networks

In an autoassociative network each pattern presented to the network serves as both the input and the output pattern. Such networks consist of mapping, middle and de-mapping layer. Middle layer, with each node representing some feature of the environment, is used for compression

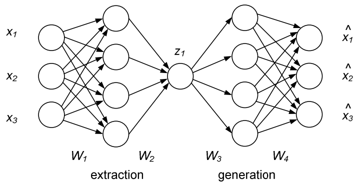

Fig. 11. Autoassociative network

or dimensionality reduction purposes, Fig. 11, (variant of Principal Component Analysis - PCA). During operation patterns are presented to the network and compared to network outputs (ideally inputs and outputs should be the same). If distance between network's inputs and outputs is to large, pattern is considered as novelty, [11, 12].

- Adaptive resonance theory (ART) network

ART is a family of different neural architectures. The first and most basic architecture is ART1 that can learn and recognize binary patterns. ART2 is a class of architectures categorizing arbitrary sequences of analog input patterns. Adaptive resonance theory network typically consists of a comparison field and a recognition field composed of neurons, a vigilance parameter, and a reset module, Fig. 12.

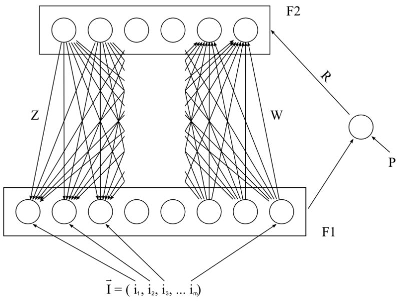

Fig. 12. Adaptive resonance theory (ART) network

When pattern is presented to the network, the network searches through categories stored for a match. If no match is found then a pattern is considered as novelty, [2,13,14].

- Radial Basis Functions (RBF) network

RBF network is a type of artificial neural network, with the radial basis functions taking on the role of the activation functions in the single hidden layer, Fig. 13. Radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, (1).

$$y(x) = \sum_{i=1}^{N} w_i \phi(\|x - c_i\|) \qquad (1)$$

Training of RBF network is accomplished in two stages: parameters of the radial basis functions are set so that they approximate model the unconditional data density of the training set, after that output weights are learned. The network outputs are estimates of Bayesian a posteriori class probabilities. Pattern with low probability is considered as a novelty, [2,4,9,15].

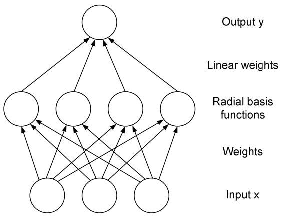

Fig. 13. RBF network

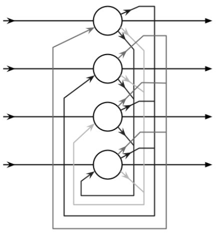

Fig. 14. Hopfield network

- Hopfield network

A Hopfield net is composed of binary threshold units with recurrent connections between them, as illustrated in Fig. 14. If the connections are symmetric, there is a global energy function. The binary threshold decision rule, (2) and (3) causes the network to settle to an energy minimum:

$$a_i = 1 \quad \text{if} \sum_j w_{ij} s_j > \theta_i \qquad (2)$$

$$a_i = 0 \quad \text{otherwise} \qquad (3)$$

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j + \sum_i \theta_i s_i \qquad (4)$$

Novelty detection is implemented by calculating the energy, (4), of the Hopfield net, after a pattern is shown. The value of the energy function E is usually lower for stored patterns and higher for other (novel) patterns, [2,4,9].

- Oscillatory networks

Oscillatory networks combine classical network models with phase dynamics. Information in the nervous system has often been considered as being represented by simultaneous discharge of a large set of neurons. The resonance amplification of network activity is used as a recognition principle for familiar stimuli. Network reaches equilibrium state in period after learning. After exposing network to a novel pattern it takes a longer time for network to reach equilibrium state compared to situation where pattern has been learned before, [2,16,17].

- Bayesian network based

A Bayesian network is a graphical model that encodes the joint probability distribution for a set of random variables. The training phase creates a tree type structure where all child nodes are the variables which feed a value to one root node for aggregation and classification of the event as normal or novelty, [2,18].

3) *Support Vector Machines (SVM)*

One approach with SVM assumes that normal points belong to high density data regions and novel patterns to low density data regions. A sphere, Fig. 15, is found that encompasses almost all points in the data set with the minimum radius. Separation of non-spherical distributed data is done in high dimensional feature space into which vectors are mapped using non-linear mapping (kernel). Second approach with SVM assumes existence of labeled data and use standard SVM for classification, [4,9,19].

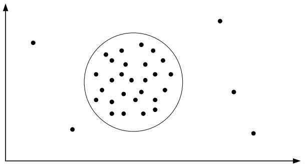

Fig. 15. Data separation using sphere in Support Vector Machines

B. Nearest Neighbor Based

This approach for novelty detection is based on the assumption that normal points have close neighbors while novelty points are located far from other points, [2,20]. In k-NN algorithm an object is classified by a majority vote of its neighbors, to the class most common amongst its *k* nearest neighbors. Two main variants of this approach are:

1) Distance based approach

Point is considered a novelty if distance to k-nn neighbor exceeds the predefined threshold

2) Density based approach

If dataset is sparse then pattern can be further away before considered as novelty, Fig. 16. This is achieved by computing local outlier factor (*LOF*). *LOF* gives an indication of how strongly an instance can be considered an outlier (novelty), [1,2].

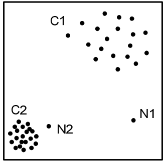

Fig 16. Density based k-NN using LOF

C. Clustering Based

This technique assume that normal data belong to large and dense clusters, while novel data don't belong to any cluster or form very small cluster. For each data point can be assigned degree of membership to each of clusters. Cluster membership is determined by comparing degree of membership to a threshold. Novelty pattern is a sample that doesn't belong to any of available clusters, [21].

D. Statistical approach

In statistical approaches stohastic distribution is used to model the data, [1,2,3].

1) Parametric approach

This approach assumes that normal data is generated from underlying parametric distribution, [1,2,3]. Here are listed some examples of this approach:

- simplest way assume distribution

Most common assumption behind normal data is a normal distribution. Pattern is considered a novelty when probability density function falls bellow a threshold (often associated with *3σ* distance from the mean *μ*), [2,3], Fig. 17.

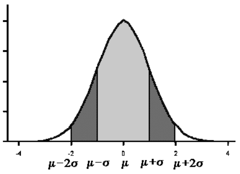

Fig. 17. Normal distribution

- probabilistic/GMM approaches

In the parametric mixture model, the component distributions are from a parametric family (often Gaussian), with unknown parameters. The *k*-component Gaussian Mixture Model (GMM), Fig. 18, is a linear combination of *k*-Gaussian pdfs, (5):

$$f_k(x) = \sum_{j=1}^{k} \pi_j \phi(x;\theta_j) \tag{5}$$

Gaussian Mixture Modeling (GMM) models general distribution estimating the density using fewer kernels than the number of patterns in training set, [3]. The parameters are estimated by the maximum likelihood (ML) criterion using the Expectation Maximization (EM) algorithm.

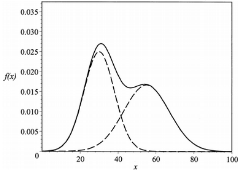

Fig. 18. Gaussian Mixture Modeling (GMM) models

- Extreme Value Theory

The extreme value theory (EVT) can be used to approach the problem of detecting novel patterns 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 [5,22].

- Hidden Markov Models (HMM)

Hidden Markov Models, Fig. 19, is statistical technique used to model sequential data (suitable for temporal patterns). HMM is comprised of a number of states together with probability of moving between pairs of states (transition probabilities), [23].

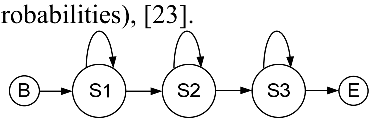

Fig. 19. Hidden Markov models (HMM)

HMM are trained on normal data (using Baum-Welch algorithm). Novel sequential pattern is detected if the probability of observing such a sequence calculated using scores below a threshold (Viterbi scoring)

- Hypothesis testing

Novelty detection can be accomplished by determining whether the test sample(s) comes from the same distribution as training data or not using t-test. If t-test shows significant difference between the two sets of measurements (normal profile and test profile), then second set is considered to contain novel patterns, [3].

Parametric approaches for estimating the probability density function have sometimes limited use since they require extensive a prior knowledge of the problem.

2) Nonparametric approach

In this approach no assumption is made about the statistical distribution of the data and is therefore more flexible, [2,3].

- Histograms

Histogram is a graphical display of tabulated frequencies, shown as bars, Fig. 20.

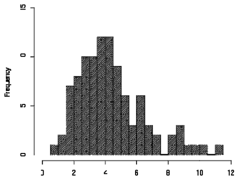

Fig. 20. Histogram - tabulated frequencies

During normal operation of a system (machine etc.) histogram acquired from collected data maintains profile of normal data. The algorithm typically defines a distance measure between a new test instance and the histogram based profile to determine if it is outlier or not, [2].

- Parzen density estimation

The Parzen window kernel density estimation can be used as a model of normality, [1,3,5]. It is essentially a data-interpolation technique. If $x_1, x_2, ..., x_n \sim f$ is an independent and identically-distributed sample of a random variable, then the kernel density approximation of its probability density function is

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad (6)$$

where K is some kernel and h is a smoothing parameter called the bandwidth, (6)**.** Often K is taken to be a standard Gaussian function with mean zero and variance 1.

- String matching approaches

These methods are based on ideas from syntactical pattern recognition that can be used instead of statistical pattern recognition if there is clear structure in the patterns [1,3,7]. Training and test data are represented by strings of symbols (possible vector quantization of input space). Novelty detection is accomplished by computing measure of dissimilarity between training and test string. Some approaches are motivated by ideas from immunology.

E. Others

1) Information theory based

This method computes information content in data using information theoretic measures, e.g. entropy, relative entropy. The basic idea is that outliers significantly alter the information content in a dataset, [1,2]. Following information theoretic measures are most commonly used in novelty detection:

- Kolmogorov complexity

Kolomogorov complexity of an object is a measure of the computational resources needed to specify the object. The more the target K varies from the baseline/normal value, the greater the likelihood the pattern is novelty.

- Entropy

Entropy is a measure of the uncertainty associated with a random variable. Any deviation from achieved entropy indicates potential intrusion.

2) Spectral decomposition based

This technique is based on eigen decomposition of data.

- Principal Component Analysis (PCA)

PCA is a statistical technique for extracting structure from dataset by performing an orthogonal basis transformation to the coordinate system in which the data is described. PCA can be used for detecting novel patterns that are orthogonal to general distribution of data. Novel patterns can be detected by looking at the last few principal components (e.g. sum of squares of the values of the last few principal components), [9].

3) Visualization based

Visualization techniques allow exploration of the structure of the data set by mapping high-dimensional data into lower dimensionality (2 or 3-D) space suitable for mapping to corresponding single point in this visualization space, [1,2]. These techniques has important role in developing good models for data, particularly when the quantity of data is large.

IV. Related Concepts in Novelty Detection

Some important concepts that are related to design of novelty detection systems are highlighted bellow, [1,2,7,9]:

A. Mode of supervision (Learning paradigms)

According to learning paradigm, novelty detection use:

1) *Supervised novelty detection* - assumes the availability of labeled training data set for normal and outlier class.
2) *Semi-Supervised novelty detection* - assumes availability of labeled instances for only one class
3) *Unsupervised novelty detection* - makes no assumption about the availability of labeled training data

*B. Output of novelty detectio*n

Typically novelty detection techniques fall into one of following two categories [2]:

1) *Labeling techniques* - assign a label (normal or outlier) to each test instance.
2) *Scoring techniques* - assign the novelty score to each pattern depending on the degree to which that pattern is considered a novelty. This introduces classifier confidence to the decision of novelty detection system.

*C. Online and Offline Activi*ty

Novelty detection methods may be implemented, [1]:

1) *Off-line* **-** batch processing, e.g. flight by flight analysis
2) *On-line* **-** real time operation - The normal behavior of a system is changing over the time. The need arises to dynamically update normal profile of a system.

D. Accuracy

Simple accuracy of the novelty detection classifier is not appropriate measure (trivial classifier that labels everything as normal would score high). Results can be better expressed in a confusion matrix (Table I) or ROC.

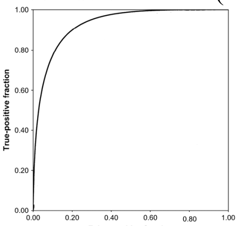

Fig. 21. Relative Operating Characteristics (ROC) curve

Table I Confusion matrix

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	*TN*	*FP*
	C	*FN*	*TP*

Normal class - NC Novelty - C

$$TPR = TP / P = TP / (TP + FN) \quad (7)$$

$$FPR = FP / N = FP / (FP + TN) \quad (8)$$

Results of classification:

TN - true negative	TP - true positive
FN - false negative	FP - false positive

The Relative Operating Characteristics (ROC), Fig. 21, can be represented by plotting the fraction of true positives (*TPR* = true positive rate) vs. the fraction of false positives (*FPR* = false positive rate), as defined in (7) and (8).

V. APPLICATIONS

Novelty detection has many practical applications in different domains and is of crucial importance in all industries, administrative applications, banking and finance systems. It includes detection of faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems etc. Here are listed some examples:

- Intelligent monitoring of high integrity systems – includes Safety Critical Systems (SCS) - health, aerospace, automotive, railway and marine systems, power generation, medical technology and Mission Critical Systems (MCS) - a high criticality with respect to the functioning of an organization, [22]
- IT Security - Network intrusion detection, Insurance / Credit card fraud detection, Mobile phone fraud detection
- Healthcare Informatics / Medical diagnostics
- Condition monitoring - decrease unscheduled outages and optimize machine performance while reducing maintenance and repair costs: Industrial Damage Detection and Engine Health Monitoring, [21]
- Image Processing / Video surveillance
- Mobile robotics
- Novel Topic Detection in Text Mining

VI. CONCLUSION

Novelty detection is a paradigm in which model of normality is constructed from normal system data. The primary objective of novelty detection is to examine if a system significantly deviates from the initial baseline condition of the system. Novelty detection methods are particularly suited for applications where most data is available from normal system operation and failures are rare. They can solve problems when novel data patterns are rare (or in some cases completely absent) during training. There exist wide range of novelty detection methods and some have been mentioned in this short review. Choice of appropriate method depends on type of input features (continuous or symbolic data), availability of labeled training data, knowledge of application domain, underlying probability distribution and available computation power if real time implementation is needed.

REFERENCES

1. A. Banerjee, V. Chandola, V. Kumar and A. Lazarević, "Anomaly Detection: A Tutorial", *Proc. of SIAM Data Mining Conference*, Atlanta, GA, April 2008,
2. V. Chandola, A. Banerjee and V. Kumar, "Outlier Detection: A Survey", Univ. of Minnesota TR 07-017, August, 2007
3. M.Markou and S.Singh, "Novelty Detection: A Review Part I: statistical approaches",*Sig. Proc.*,Vol. 83,No.12,Dec. 2003
4. M.Markou and S.Singh,"Novelty Detection:A Review Part II: neural network approaches",*Sig.Proc*,Vol.83,No.12,Dec.2003
5. L. Tarassenko, D. A. Clifton , P. R. Bannister, S. King and D. King, "Novelty Detection", in Worden, K., et al. (eds): *Encyclopaedia of Structural Health Monitoring*, Wiley, 2009
6. J. S. Anstey, D. K. Peters and C. Dawson, "Discovering Novelty in Time Series Data", *Proc. Newfoundland Electrical and Computer Engineering Conference*, *IEEE*, Newfoundland and Labrador Section, November 2005
7. R. Schalkoff, *Pattern Recognition, Statistical, Structural and Neural Approach*, John Wiley, New York, NY, USA, 1991
8. T.Kohonen, "Self-Organizing Maps",2nd.Ed.,Springer, 1997
9. S. Marsland, Novety Detection in Learning Systems, *Robotics and Autonomous Systems*, 51(2-3):191-206, 2005.
10. D. Martinez, Neural tree density estimation for novelty detection, I*EEE Transactions on Neural Networks*, Volume: 9 Issue: 2, pp. 330 - 338, Mar 1998
11. H. Sohn, K. Worden and C. R. Farrar, Novelty Detection Using Auto-Associative Neural Network, *Proc. Symp. Identification of Mechanical Systems: Int'l Mechanical Eng. Congress and Exposition*, New York, USA, November 2001
12. H. Lee, B. Hwang and S. Cho, Analysis of Novelty Detection Properties of Autoassociative MLP, *Journal of the Korean Institute of Industrial Engineering*,Vol.28,No2, 2002
13. B. Rowland and A. S. Maida, "Spatiotemporal Novelty Detection Using Resonance Networks", *Proceedings of the 17th Annual Florida AI Research Society Conference* pp. 676-681, Miami Beach, FL, May 2004,
14. T.Tanaka and A.Weitzenfeld,"Adaptive Resonance Theory", in *Neural Simulation Language*, The MIT Press, 2002
15. A.L.I. Oliveira1, F.. Neto and S. R.L. Meira, "Combining MLP and RBF Neural Networks for Novelty Detection in Short Time Series", *Proc. of MICAI 2004*, Mexico, 2004
16. R. Borisyuk, M. Denham, F. Hoppensteadt, Y. Kazanovich and Olga Vinogradova, "Oscillatory Model of Novelty Detection", *Network Computation in Neural Systems*, Vol. 12, No 1, pages 1 - 20, Jan. 2001
17. T. V. Ho and J. Rouat, "A Novelty Detector Using a Network of Integrate and Fire Neurons", *Proceedings of the 7th ICANN,* Lausanne, Switzerland, 1997
18. Z. Ghahramani, "Learning Dynamic Bayesian Networks", *Lecture Notes In Computer Science*, Vol. 1387, pp. 168-197, Springer, 1997
19. B.Schölkopf, R.Williamson, A.Smolax, J.Shawe-Taylory and J. Platt, "Support Vector Method for Novelty Detection", in *Advances in Neural Information Processing Systems 12*, The MIT Press, June 2000
20. V. Hautamäki, "Outlier Detection Using k-Nearest Neighbour Graph", *17th International Conference on Pattern Recognition (ICPR'04)* - Vol. 3, Cambridge,UK 2004
21. D. Miljković,"Novelty Detection In Machine Vibration Data Based On Cluster Intraset Distance",*Proc. CTS,* MIPRO 2008
22. I.S. Sundaram,I.G.D.Strachan,D.A.Clifton,L.Tarassenko and S.King, "Aircraft Engine Health Monitoring using Density Modelling and Extreme Value Statistics", *Proc. 6th Intern. Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, Dublin, Ireland, June 2009
23. J. R. Noriss, "Markov Chains",Cambridge Univ. Press., 1997
24. D.A.Clifton,P.R. Bannister,L.Tarassenko and S. King, "High Dimensional Visualisation for Novelty Detection", *Proc. of 5th Int.Conf. on Condition Monitoring*,Edinburgh,U.K., 2008