[1. Outlier Detection 2](#_Toc337374098)

[1.1. Classification of Outlier detection problems 2](#_Toc337374099)

[1.1.1. Data characteristics 2](#_Toc337374100)

[1.1.1.1. Input Data 3](#_Toc337374101)

[1.1.1.1.1. Dataset characteristics 3](#_Toc337374102)

[1.1.1.1.2. Type of supervision 3](#_Toc337374103)

[1.1.1.1.3. Outlier Types 4](#_Toc337374104)

[1.1.1.2. Output data 5](#_Toc337374105)

[1.1.2. Knowledge Disciplines 5](#_Toc337374106)

[1.1.3. Application Domains 5](#_Toc337374107)

[2. Condition Monitoring 6](#_Toc337374108)

[2.1. Data Acquisition 7](#_Toc337374109)

[2.2. Data Analysis 7](#_Toc337374110)

[2.2.1. Time Domain Data 8](#_Toc337374111)

[2.2.2. Modal Domain Data 8](#_Toc337374112)

[2.2.3. Frequency Domain 8](#_Toc337374113)

[2.2.4. Time Frequency Domain 9](#_Toc337374114)

[2.3. Feature Selection 9](#_Toc337374115)

[2.4. Decision Techniques 10](#_Toc337374116)

[2.4.1. Classification based approaches 10](#_Toc337374117)

[2.4.1.1. Neural Networks 11](#_Toc337374118)

[2.4.1.2. Support Vector Machines 12](#_Toc337374119)

[2.4.1.3. Rule Based Approaches 14](#_Toc337374120)

[2.4.2. Statistical Methods 15](#_Toc337374121)

[2.4.2.1. Parametric Approaches 15](#_Toc337374122)

[2.4.2.1.1. Gaussian Models 15](#_Toc337374123)

[2.4.2.1.2. Regression Models 17](#_Toc337374124)

[2.4.2.1.3. Mixture of parametric models 18](#_Toc337374125)

[2.4.2.1.4. Markov and Hidden Markov Models 18](#_Toc337374126)

[2.4.2.2. Non-Parametric Approaches 20](#_Toc337374127)

[2.4.3. Spectral Decomposition 20](#_Toc337374128)

[Literaturverzeichnis 21](#_Toc337374129)

# Outlier Detection

Outlier detection is the problem of finding elements in datasets which are unusual or conspicuous compared to other elements in the same dataset. A classical definition was given by Grubbs (Grubbs, 1969):

“An outlying observation, or outlier, is one that appears to deviate markedly from other members in the same sample in which it occurs”.

Outliers can occur in many different application domains and depending on the contexts, they are sometimes also referred to as anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants (Banerjee, Chandola, & Kumar, 2007).

The various techniques which can be applied to detect outliers in datasets involve methods from different knowledge disciplines such as statistics, machine learning and pattern recognition.

## Classification of Outlier detection problems

The choice of methods for an outlier detection problem strongly depends on the characteristics of a given dataset and the application domain. Thus, a preliminary and fundamental step in any outlier detection scenario is the classification of the problem with respect to dataset characteristics and the application domain constraints. Chandola et al. provided a system for classification of outlier detection problems in (Banerjee, Chandola, & Kumar, 2007), by identifying the three key dimensions Data, Knowledge Disciplines and Application Domain. And describing the sub-aspects of each dimension.

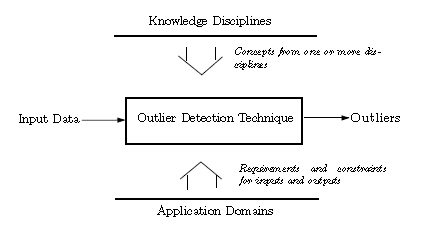


Fig. 1 Classification of Outlier Detection Problems

### Data characteristics

The data dimension comprises characteristics of the input dataset as well as requirements regarding the expected output of an outlier detection technique.

#### Input Data

The input data can be characterized by the following aspects:

* Dataset characteristics
* Type of Supervision
* Type of Outlier

##### Dataset characteristics

The concept of data can be separated into a formal description of the data describing the general structure of data points, and individual data points representing real world data.

The structure of data points is defined by one (in the case of univariate data) or multiple attributes (in the case of multivariate data). Attributes have data types which can be either continuous numerical, discrete numerical or nominal.

The actual data points are represented by data instances or data objects, with individual values for each attribute.

Collections of such data instances are the input of outlier detection sequences. In some cases, data has to be preprocessed to reduce the number of attributes (2.3) or to complete missing data values.

##### Type of supervision

Data instances can contain additional information in form of a special attribute, generally known as target attribute. Just as the other attributes, a target value has a type which is either numerical or nominal. Generally speaking, the target attribute represents a mapping from the input space to a target space, which assigns a value to a data point. Target attributes in outlier detection data are usually nominal values, classifying a data point as either normal or outlier.

Some Outlier Detection techniques use data instances with target values as training data to build predictive models. *Supervised techniques* use data instances from both the outlier and the normal class as training data to create separate models for both classes. New data instances are compared to both models and assigned to the model-class with the closest match.

Due to the specific nature of outlier detection problems, it can be prohibitively expensive to obtain sample data for both the normal and the outlier class. In many cases it is much more difficult to get training data for the outlier class, since it would involve the arrangement of accidents (e.g. in air craft condition monitoring), the destructive manipulation of expensive machinery or the simulation of a great number of abnormal patterns (e.g. in network intrusion detection). Techniques that assume the availability of training data instances for only one class are known as *Semisupervised Techniques.* Such techniques build a model that matches the available class data (usually the normal class), and assign any data instance significantly deviating from the model to the outlier class.

The third category of outlier detection techniques, *Unsupervised Techniques,* do not assume the availability of any training data and are thus the most generally applicable techniques. Methods in this category usually use some kind of statistical evidence derived from the data. One trivial statistical assumption can be based on the observation, that normal data points are far more frequent than outlier data points in data sets. Clustering algorithms make use of this observation, by assigning rare points or patterns to the outlier class.

##### Outlier Types

Chandola et al. distinguish three Outlier Types with respect to the composition of an outlier and its relation to the rest of the data (Banerjee, Chandola, & Kumar, 2007). The identification of the type of outliers in a given data set is an important preliminary step in any outlier detection scenario, since it determines and limits the selection of techniques.

Type I outliers

Type I outliers are individual data points with attribute values that are inconsistent with the attribute values of normal data points. Fig 2 shows an example with 5 individual outliers.

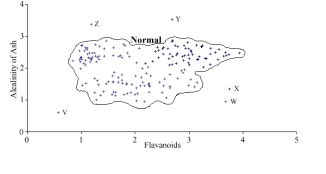


Fig. 2 Dataset with 5 individual outliers (Hodge & Austin, 2004)

Type II outliers

Type II outliers are like type I individual data instances, but are outliers only within a certain context and considered as normal when they occur in other contexts. The type of context depends on the dataset and the application domain and can be either spatial or sequential. In the temperature dataset in Fig. 2, t2 is a type II outlier in the context of “summer”, whereas the same temperature is considered normal in the context of “winter”.

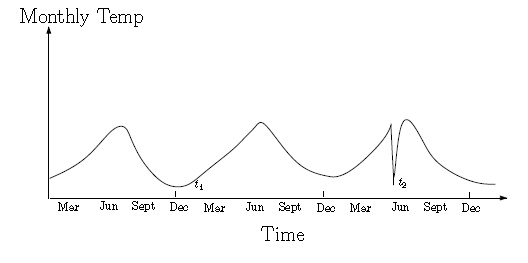


Fig. 3 Type II outlier t2 in a temperature time series (Banerjee, Chandola, & Kumar, 2007)

Type II outliers can be seen as an extension of type I outliers, with additional information specifying contextual information for each data point. Thus, data instances of type II outliers need an additional attribute set describing the context in which the individual data point occurred.

Type III outliers

Other than type I and type II outliers, type III outliers are not individual data points, but subsets of data. In other words, type 3 outliers are not outliers due to their individual attribute values, but due to their occurrence together with other data points in a pattern or at a place which is considered anomalous. This implies that such outliers only occur in datasets of spatial or sequential nature.

In the electrocardiogram data in Fig. 4, the long sequence of certain signal level signifies an outlier event.

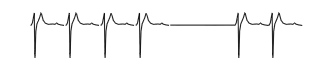


Fig. 4 Type III outlier in an human electrocardiogram output (Banerjee, Chandola, & Kumar, 2007)

#### Output data

The output of an outlier detection sequence can simply be the input data instances, extended by an additional attribute labeling the data instances as either normal or output data. Techniques which label data in such a binary manner are known as *Labeling Techniques* or *Classification Algorithms.* Depending on the application domain, a more differenciated qualification of data is needed. *Scoring Techniques* assign a score to data instances, indicating to which degree a data instance is considered an outlier or normal data. The output of such techniques usually consists of a ranked list, allowing a further examination by experts or algorithms.

### Knowledge Disciplines

The techniques used for detecting and handling outliers are various and depend on the characteristics of the outlier data, the expected output format and other constraints. The vast majority of approaches can be assigned to one or more of the following categories:

* Classification Based Approaches (such as Neural Networks, Rule Based Models etc.)
* Clustering Based Approaches
* Nearest Neighbor based approaches
* Statistical Approaches
* Information Theory based approaches
* Spectral Decomposition based approaches
* Visualization based approaches

### Application Domains

Some popular applications for outlier detection are:

* Fraud Detection – detection of fraudulent application for credit cards, mobile, insurance etc.
* Loan application
* Intrusion detection – detecting unauthorized access in computer networks
* Activity Monitoring
* Network performance
* condition monitoring – Fault detection, structural damage detection, time series monitoring
* Structural defect detection
* Medical and Health data
* Condition monitoring

# Condition Monitoring

Condition monitoring is the automatic supervision of industrial machinery and structures, with the aim to prevent and detect damages. It can be defined as “a process of monitoring a system by studying certain selected parameters in such a way that significant changes of those parameters are related to a developing failure”. (Marwala, 2012)

The two main application areas in Condition monitoring are Fault Detection in Mechanical Units (Motors, Turbines, pipelines etc.) and Structural Defect Detection in industrially produced units such as beams, airframes etc.

In both areas, data is usually acquired by sensors monitoring relevant parameters of the particular unit or process, which often involve some kind of vibrational (in structures) or rotational (e.g. in motors) parameters or other time-dependent factors such as the temperature over time. Outliers in Condition Monitoring mostly occur in a certain context or as a sequence of events and can thus be classified as either Type II or Type III outliers.

Condition Monitoring scenarios generally comprise several consecutive stages, illustrated in fig. 1.

Data acquisition

Data analysis

Feature selection

Decision

Condition Diagnosis

Fig. 5 Condition Monitoring stages

The first step consists of data acquisition by sensors such as thermometers, strain gauges or accelerometers which are directly attached to the monitored process and collect the raw data.

In the data analysis stage, time domain raw data can be transformed by techniques revealing the relevant information contained in raw data. Popular techniques include the Fourier Transformation, Modal Analysis or wavelets.

Datasets generated in the first or second stages can contain a large number of attributes, where only some are good indicators for potential faults and others are redundant or irrelevant. Such datasets can in a third step be processed by feature selection methods that identify the relevant aspects of the data and discard the others. This is often achieved by applying *Principal Component Analysis* or *Independent Component Analysis.*

In the Fourth step, datasets are analysed by knowledge establishing techniques which interpret the data and classify data instances as normal or abnormal. Popular techniques used in Condition Monitoring can be assigned to the categories non-parametric / parametric statistical approaches, Neural Nets and Rule Based approaches.

The final step is the Condition Diagnosis based on the results produced by step 4. This includes the assessment of potential faults with respect to their location, severance, their possible impact on the system and lifetime estimation as illustrated in fig. ..

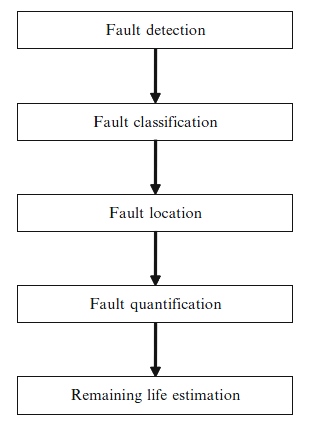


Fig. 6 Fault Assessment

Condition Diagnosis requires a thorough understanding of the application domain and the particular system and is usually done by domain experts.

Of the stages illustrated in Fig. 1, the 1st,4th and 5th steps are mandatory, while the 2nd and 3rd are mandatory and only applied if required by dataset characteristics or other constraints.

The Following paragraphs describe the techniques applied in the stages of a Condition Monitoring scenario in more detail.

## Data Acquisition

## Data Analysis

Condition Monitoring data can be represented in four main domains: *Time Domain*, *Modal Domain*, *Frequency Domain* and *Time-Frequency* Domain.

### Time Domain Data

Data acquired in the data acquisition phase is *Time-Domain* data, with sensor signals measured over historical time. *Time Domain* data can be directly used with statistical techniques which base their decision on statistical characteristics such as the mean or the variance of the data. Other techniques in decision stage require some sort of data transformation from the *Time Domain* into one of the other three domains, which is done in the data analysis stage.

### Modal Domain Data

Modal Domain data represents physical properties of structures such as natural frequencies, the damping ratio and modal shapes. Techniques analysing modal domain data are based on the observation that these physical properties change in the presence of faults somewhere in the structure.

The most widely used technique to extract modal properties from raw data *Modal Analysis* (Ewin), which is based on the general equation for damped harmonic oscillators given by

(9.1)

Where M, C and K are matrices representing the distributed mass, damping and stiffness of a structure, x is the displacement vector and F is a matric representing the applied force.

The transformation from the time domain into the modal domain is done by forming an eigenvalue expression for the mode of 9.1, given by

Where is the complex eigenvalue with its imaginary part corresponding to the natural frequency and is the complex mode shape vector corresponding to the normalized mode shape .

### Frequency Domain

The Frequencies contained in time domain raw data can be extracted by using the *Fourier Transform*. In practice, usually the computationally efficient *Fast Fourier Transform* is used, which is given by

(9.2.1)

If the Fourier transform of the excitation of a structure is given by F and the Fourier transform of the according response is given by X, the matrix of the *Frequency Response Functions (FRF)* is the ratio

Where is the response at location i to an excitation F at location j.

### Time Frequency Domain

Modal Analysis and Fourier Transform can only deal with stationary frequencies. However, in some scenarios the detection of certain types of faults requires a measurement of frequency changes over time. Examples for such non-stationary scenarios are noise analysis, or the acceleration of a starting train. Techniques that transform raw data into the Time-Frequency Domain include the *Short Time Fourier Transform (STFT)*, the *Wavelet Transform (WT)*, and the *Wigner-Ville-Distribution (WVD)*.

In the *STFT* a Time Frequency Transformation is accomplished by subdividing the time space into small windows of a certain time width and successively calculating the Fourier Transform for each of these time intervals. By doing so, a time-frequency is obtained. The time-frequency correspondents in are calculated by

Which basically is the Fourier Transform multiplied with the complex conjugate of a window function .

Since the Frequency resolution of a Fourier transform is proportional to the time interval over which it is taken, smaller windows result in a lower frequency resolution. Larger time windows increase the frequency-domain resolution but at the same time decrease the time-domain resolution. The *Continuous Wavelet Transformation CWT* was developed to overcome this uncertainty relation between time and frequency solution. For a given point at the time axis and a scale the *CWT* of a time signal x(t) can be calculated by

## Feature Selection

Multivariate data can contain a lot of information which is irrelevant for the solution of a given problem. At the same time, the computational cost of many knowledge generation algorithms increases considerably with the data dimensionality. This complexity problem, also known as *the curse of dimensionality* can be tackled by techniques that try to reduce the dimensionality of data by identifying the relevant features in datasets. The most widely used *Feature Selection* technique is the *Principal Component Analysis (PCA).*

*PCA* achieves a reduction of data complexity by selecting the most relevant data dimensions and transforming the data to a new coordinate system spanned by these dimensions. The Principal Components of a dataset are identified by finding the Eigenvalues and the corresponding Eigenvectors of the Covariance Matrix which is given by

(2.3.1)

where is the sample set mean

(2.3.2)

And is a data instance.

The Eigenvalues and Eigenvectors are then calculated by solving

(2.3.3)

where is the i-th eigenvector corresponding to the i-th eigenvalue .

The data is finally compressed by creating a transformation matrix with the eigenvectors of the n largest eigenvalues and transforming the dataset to the reduced coordinate system. Fig. 7 illustrates the reduction of a 2-dimensional data set to 1 dimension, without loss of the significant variance information contained in the original dataset.

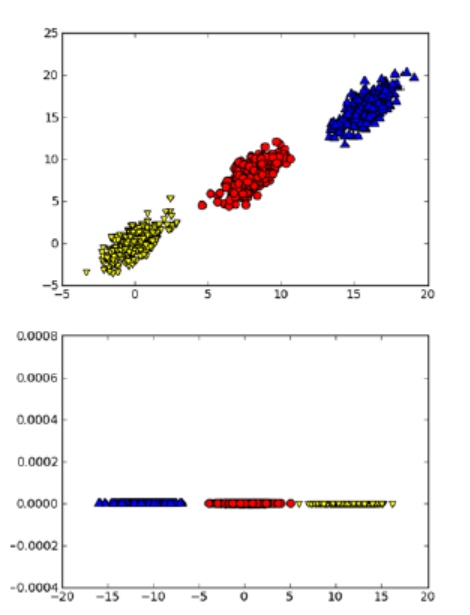


Fig. 7 PCA-illustration

## Decision Techniques

### Classification based approaches

Classification techniques consist of two phases, a learning phase and a testing phase. In the learning phase, a classification model is built by learning from training data. In the testing phase, unseen and unlabeled data instances are labeled according to the classification models. The availability of labeled training data is a fundamental requirement for the application of classification based approaches. Depending on the type of supervision, models are built for both the normal and the outlier class (*supervised)* or for only one class (*semi-supervised)*, which is usually the normal class. In the first case, the assignment of a class label for an individual data instance is based on a comparison of the instance with both models. In the latter case, a data instance is compared with the model for just one class and either accepted as belonging to this class, or rejected if the instance deviates significantly from the learnt model.

#### Neural Networks

“The term ‘neural network’ has its origins in attempts to find mathematical representations of information processing in biological systems” (Bishop C. M., 2009). A large number of variations exist, many of which are based on or in some way enhance the basic concept of *Multilayer Perceptrons* (Bishop C. M., 2009)*,* (Banerjee, Chandola, & Kumar, 2007). *Multilayer Perceptrons* consist of a number of interconnected nodes, which are structured in an input layer, one or several hidden layers and an output layer.

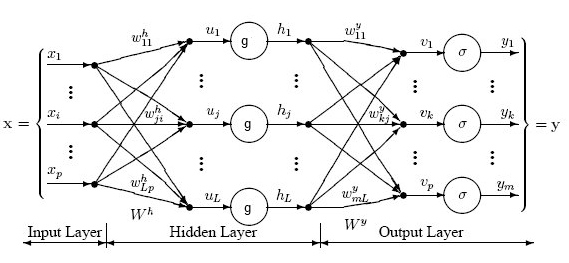


Fig. 8 Multilayerd Perceptron

The input to a hidden layer node is a linear combination of the input value vector and a weight vector

(11.1.1.1)

where P is the input vector size, h indicates that the weights are the weights of the hidden layer, j specifies a hidden node and is the ith element of the input vector.

The output of a hidden node is the result of a function

(11.1.1.2)

The input to the output layer nodes is a linear combination of the hidden layer output vector and a weight vector:

(11.1.1.3)

The output of the output layer nodes is often calculated by a sigmoid function so that

(11.1.1.4)

With (1), (3) and (4) the overall output of an output layer unit gives

(11.1.1.5)

The parameters of Multilayer Perceptrons, the weight vectors , are trained during the training phase. Given a training set with target value data instances (supervised or semisupervised data), the weights can be adjusted by minimizing the error function

(11.1.1.6)

Where N is the total number of training data instances and is the label vector of the n-th training data instance. With a method called *Batch Gradient Descent* the weights can be iteratively updated by

where is known as the learning rate.

Different types of Neural Nets have been used in a wide range of industrial and medical applications. Bishop presented a general approach for the application of Neural Networks in Outlier Detection problems by using a confidence measure corresponding with the network outputs to detect novelty. The method was applied in a semisupervised setting, where the network was trained to classify a number of multiphase profiles of oil gas and water in oil pipes. Input data that generated outputs with a confidence level below a certain threshold was considered as novelty data, not belonging to any of the trained classes (Bishop C. , 1994).

#### Support Vector Machines

In a 2-dimensional classification problem with 2 classes, where the training data points are linearly separable, a simple line can separate the points of class 1 from the points of class 2, as illustrated in Fig 9.

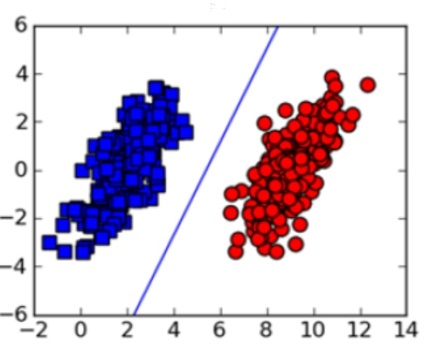


Fig. 9 Linearly separable classes

Generalized to D dimensional data, linearly separable classes can be separated by a hyperplane of the form

(2.4.1.2.1)

where is a vector of adaptable model parameters.

*Support Vector Machines* essentially optimize the position of a separating hyperplane by maximizing the distance between the points closest do the decision boundary (i.e. the hyperplane) and the decision boundary. This distance is also known as *Margin* as shown in Fig…

*<illustration aus bishop…>*

Fig. shows that the points closest to the hyperplane, also known as *Support Vectors,* determine the location of the decision boundary.

The calculation of the margin includes the computation of the perpendicular distance between a point and a hyperplane as defined above, which is given by

(2.4.1.2.2)

Given a training set with target values where and with a choice of the hyperplane parameters and b such that for points having as target value and for points with target value , for correctly classified classified points always holds

(2.4.1.2.3)

So that equation 2.4.1.2.2 can be expressed by

(2.4.1.2.4)

With the margin given by the perpendicular distance to the closest point in the training set, the parameters and b have to be optimized in order to maximize the margin, which can be done by solving

(2.4.1.2.5)

This very complex optimization problem is converted to an equivalent, but much easier problem by setting

(2.4.1.2.6)

for the support vectors, in which case all data points will satisfy the constraints

(2.4.1.2.7)

The optimization problem is now reduced to the maximization of which under the constraint given by equation 2.4.1.2.7. With the help of *Lagrange Multipliers*, the maximization problem can be reformulated as

…

#### Rule Based Approaches

Rule Based techniques generate sets of rules capturing the normal behavior of a system. An instance that doesn’t fit any of these rules is considered an outlier. This category comprises several different approaches and covers both supervised and unsupervised techniques. A big advantage of Rule Based Approaches is that the created models often consist of logical expressions such as “if-then” statements which makes them interpretable by humans. They are therefore often applied in areas where a certain domain specific expert knowledge is involved in creating the models and interpreting the results.

One very popular technique is *Association Rule Mining,* where a set of rules is created based on transactions. A transaction T is a subset , where is a finite item set . An Association Rules is an implication of the form , where and . Association rules are found by iterating over a set of Transactions in a database and accumulating the frequency of implications. There are two basic measurements, support s and confidence c. The support of an association rule is defined as the ratio of transactions in a database that contain to the complete number of transactions in the database. Confidence of an association rule is the ratio between the number of transactions containing to the number of transactions containing X. An association rule is added to the model if its support and confidence are above a certain user defined threshold.

Yairi et al used an Association Rule Mining approach for online creation of a set of association rules used to monitor spaceship condition (Yairi, Kato, & Hori, 2001).

### Statistical Methods

From a statistical point of view, an outlier can be considered as “an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed” (Anscombe & Guttman, 1960). Statistical methods try to establish a statistical model for the distribution of a given dataset and then estimate the probability for data instances of being generated by that distribution. If the probability is beneath a certain threshold, a data instance is labeled as an outlier. Just as classification methods, statistical methods consist of two phases: The training phase where the statistical model is generated and the test phase in which the outlier probability for unseen data is estimated.

Statistical techniques can generally be grouped in the two categories *Parametric Techniques* and *Non-Parametric Techniques*, with different approaches in training and test phase.

*Parametric Statistical Approaches* are based on a prior assumption of the data distribution and estimate the values of the distribution parameters during training phase. Since single distributions often don’t fit real world data, Parametric Statistical approaches can combine several distributions to approximate the data. In test phase, parametric statistical techniques often use some variation of distance measurement between a data instance and the estimated mean of the distribution and designate data points as outliers, if that distance is above a certain threshold.

Non-parametric approaches make no assumption whatsoever about the underlying distribution. Some common techniques create *Histograms* during training phase, where the frequency of occurrence of data instances is simply counted. Others try to estimate the probability density of a dataset with unknown distribution. In test phase, typically the distance between a data instance and the statistical model is measured. If the model is a histogram, the distance between a data instance and the histogram categories is measured and the data instance is assigned to the closest category.

#### Parametric Approaches

##### Gaussian Models

Gaussian models assume a Gaussian distribution as statistical model. Due to its prominence in statistics, many parametric statistical approaches are based on the Gaussian distribution which for univariate data can be written in the form

(12.1.1.1)

where is the mean and is the variance.

In case of multivariate data, (1) takes on the form

(12.1.1.2)

where D denotes the dimension of a data instance vector , is a D-dimensional mean vector, is the covariance matrix, is the inverse and is the determinant of the covariance matrix.

The Training phase usually includes the estimation of the parameters () and () by *Maximum Likelihood Estimation*, which maximizes the probability that a given dataset is generated by a set of parameters

(12.1.1.3)

Where X is a matrix of datasets and is a vector of statistical parameters.

The probability of a Gaussian distribution given the mean and the variance is the joint probability of all data instances given the two parameters:

(12.1.1.4)

Where X is the complete dataset and N is the number of data instances in the dataset.

Maximizing (12.1.1.4) with respect to the maximum likelihood for the mean gives

(12.1.1.5)

and

(12.1.1.6)

Maximizing (12.1.1.4) with respect to the maximum likelihood for the variance gives

(12.1.1.6)

for univariate distributions and

(12.1.1.7)

For multivariate distributions.

In testing phase, several methods are used, of which the *Box-plot* rule and (variants of) the *Grubbs* test are very common.

A box plot (fig 6) is a graph displaying the shape of a distribution, with its central value its spread and the maximum and minimum values.

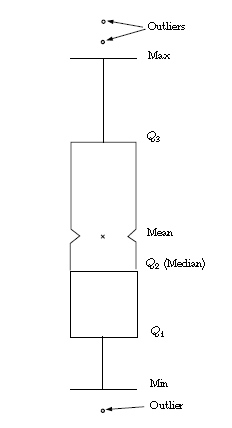


Fig. 10 Box Plot

Any data point lying beyond the extreme values is treated as outlier. Analysis of Box Plot graphs can be used for univariate and multivariate data and where applied in a number of different application domains (Banerjee, Chandola, & Kumar, 2007).

The standard *Grubb’s* test can be applied to univariate datasets. It iteratively deletes outliers with the highest value

where and s being the mean of the dataset . Several newer approaches can be used to test multivariate data as well.

##### Regression Models

The training phase of all regression techniques consists in fitting some form of regression function on the data. The general form of a regression model is given by

(12.1.2.1)

Where is the vector of (statistical) parameters and is any linear or non-linear function of the input vector , usually denoted as *basis function*.

If and M is set to the dimension D of , 12.1.2.1 takes the form of the most simple linear model for regression which is just a linear combination of the input vector and the parameter vector

(12.1.2.2)

For univariate data sets, the basis function can be set to which yields the polynomial regression function

(12.1.2.3)

Common choices of nonlinear basis functions are Gaussian functions, Sigmoid functions or wavelet functions.

The second main step in any regression technique after selecting the basis function is the calculation of the parameter vector , which is usually done by minimizing an error function like the sum-of-squares error function

(12.1.2.4)

where is the label of the n-th training data instance.

##### Mixture of parametric models

Due to the observation that single distributions often don’t fit a complete real world dataset, some approaches in the field of parametric statistical models combine several distributions. Depending on the type of supervisions, this can basically be done in two ways. In the case of supervised datasets, the normal class and the outlier class are modeled as separate statistical distributions during training phase. The test phase then involves checking which distribution a data instance most likely belongs to.

The semisupervised version of parametric mixture techniques creates several models for just one class, usually the normal class. Testing then consists of checking whether a data instance belongs to any of the models for the normal class and labeling an instance as outlier if it doesn’t.

##### Markov and Hidden Markov Models

The use of Hidden Markov Models in Pattern Recognition problems was described by Rabiner (Rabiner, 1989). A Markov model is a state model with a finite set of N states and values representing the transition probability from state to . Fig. 11 illustrates a Markov model with N=5 and non-zero transition probabilities between the 5 states.

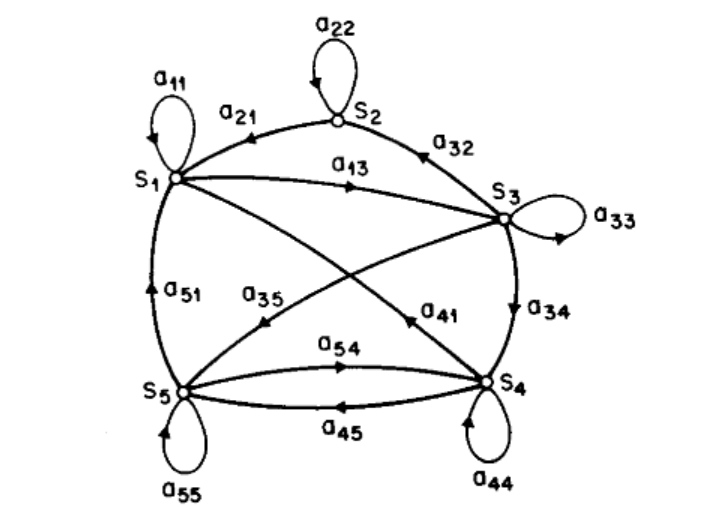


Fig. 11 Markov Model

In *Hidden Markov models*, the states of a system cannot be observed directly, but can be deduced through observations correlated with the states.

The Elements of *Hidden Markov Models are*

1. The set S with the N states
2. A set of M observations
3. The state transition probability distribution , where

i.e. the probability that State succeeds state .

1. The observation symbol probability distribution for a given state j, where

i.e. the probability of observation at time t given the state at that time.

1. The initial state distribution where

i.e. the Probability that is the initial state.

With A and B implying N,M,S and V respectively, the parameter sets of *HMM* can be written in a compact notation as

In real world application of *HMM* three different calculation scenarios can occur.

In the first scenario, an observation sequence and the model are given and the question is how to calculate the probability efficiently.

In a second scenario, an observation sequence and a model are given and the challenge is to establish the most likely State sequence corresponding with the observation sequence.

The third scenario is the training scenario, where the parameters of the model are adjusted in a way that maximizes , where O can be seen as training observation data.

#### Non-Parametric Approaches

Non-parametric approaches comprise all techniques that make no prior assumption on the data distribution. The most widely used methods in this category are approaches based on histogram analysis and *Finite State Automata*.

*Histogram* techniques are usually semi-supervised and create a profile of the normal data, based on the frequency of attribute-value occurrences in a training set. Testing involves checking, if a data instances fits the normal data profile, which is usually done by some sort of distance measurement. Building of and testing against histograms is easy in the case of univariate data, where the histogram consists of frequency bins representing the different value ranges of that single attribute in the training set. In multivariate scenarios, either a single significant feature has to be selected, or histograms have to be created and maintained for several features. In the latter case, testing involves some kind of combined distance measurement of data instances and the histograms.

*Finite State Automata* can be used to model semi-supervised data of temporal or sequential nature. The training phase consists of generating finite state machines representing the states and transitions based on the normal or outlier class data.

Another common subcategory of non-parametric approaches are techniques measuring the probability distribution of a dataset. A popular method is the *Parzen Window Estimation* which involves the use of kernel functions to estimate the density distribution of a dataset. A data instance which lies in a low area of the probability distribution function is declared an outlier.

### Spectral Decomposition

The Spectral Decomposition of a real symmetric matrix A has the form

(2.4.3.1)

Where U is a matrix whose columns are the (orthonormal) eigenvectors of A and is a diagonal matrix with the eigenvalues of A as diagonal elements. Multiplied with U, 2.4.3.1 can be written as

(2.4.3.2)

By defining the column vector of U as and the eigenvalue as , 2.4.3.2 can be re-written as

(2.4.3.3)

If A is the covariance matrix, 2.4.3.3 becomes the base formula for the *Principal Component Analysis.*

Fujimaki et al. transformed spacecraft time series data at discrete time points into a high dimensional feature space, identified the *Principal Components* at each time point and detected anomalies through significant direction changes of the *PCA* vectors (Fujimaki, Yaira, & Machida, 2005).

# Literaturverzeichnis

Anscombe, F., & Guttman, I. (1960). Rejection of outliers. *Technometrics 2*, S. 123-147.

Banerjee, A., Chandola, V., & Kumar, V. (2007). Outlier Detection: A Survey.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley & Sons.

Bishop, C. (1994). Novelty Detection an Neural Network validation. *IEE Proceedings - Vision, Image and Signal processing, Vol. 141, No. 4*, S. 217-222.

Bishop, C. M. (2009). *Pattern Recognition and Machine Learning.* Springer.

Fujimaki, R., Yaira, T., & Machida, K. (2005). An Approach to Spacecraft Anomaly Detection Problem Using Kernel Feature Space. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*.

Grubbs, F. (1969). Procedures for detecting outlying observations in samples. *Technometrics 11*, S. 1-21.

Hodge, V., & Austin, J. (1. October 2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review, Vol 22, No.2*, S. pp. 85-126.

Markou, M., & Singh, S. (2003). Novelty Detection: A Review part 2: Neural Network based approaches. *Signal Processing 83*, S. 2499-2521.

Markou, M., & Singh, S. (2003). Novelty detection: a review--part1: Statistical approaches. *Signal Processing 83*, S. 2481-2497.

Marwala, T. (2012). *Condition Monitoring Using Computational Intelligence Methods.* Springer.

Rabiner, L. R. (February 1989). A Tutorial on Hidden Markov Models. *Proceedings of the IEEE, Vol 77, No.2*.

Yairi, T., Kato, Y., & Hori, K. (2001). Fault Detection by Mining Association Rules from House-keeping Data. *Proceeding of the 6th International Symposium on Artificial Intelligence and Robotics & Automation in Space*.

[Fig. 1 Classification of Outlier Detection Problems 2](#_Toc336767952)

[Fig. 2 Dataset with 5 individual outliers (Hodge & Austin, 2004) 4](#_Toc336767953)

[Fig. 3 Type II outlier t2 in a temperature time series (Banerjee, Chandola, & Kumar, 2007) 5](#_Toc336767954)

[Fig. 4 Type III outlier in an human electrocardiogram output (Banerjee, Chandola, & Kumar, 2007) 5](#_Toc336767955)

[Fig. 5 Condition Monitoring stages 7](#_Toc336767956)

[Fig. 6 Fault Assessment 8](#_Toc336767957)

[Fig. 7 PCA-illustration 10](#_Toc336767958)

[Fig. 8 Multilayerd Perceptron 11](#_Toc336767959)

[Fig. 9 Box Plot 15](#_Toc336767960)

[Fig. 10 Markov Model 17](#_Toc336767961)