14

Outlier Detection Methods for Industrial Applications

Silvia Cateni, Valentina Colla and Marco Vannucci
Scuola Superiore Sant Anna, Pisa
Italy

1. Introduction

An outlier is an observation (or measurement) that is different with respect to the other values contained in a given dataset. Outliers can be due to several causes. The measurement can be incorrectly observed, recorded or entered into the process computer, the observed datum can come from a different population with respect to the normal situation and thus is correctly measured but represents a rare event. In literature different definitions of outlier exist: the most commonly referred are reported in the following:

- "An outlier is an observation that deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism " (Hawkins, 1980).
- "An outlier is an observation (or subset of observations) which appear to be inconsistent with the remainder of the dataset" (Barnet & Lewis, 1994).
- "An outlier is an observation that lies outside the overall pattern of a distribution" (Moore and McCabe, 1999).
- "Outliers are those data records that do not follow any pattern in an application" (Chen and al., 2002).
- "An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data" (Ramasmawy at al., 2000).

Many data mining algorithms try to minimize the influence of outliers for instance on a final model to develop, or to eliminate them in the data pre-processing phase. However, a data miner should be careful when automatically detecting and eliminating outliers because, if the data are correct, their elimination can cause the loss of important hidden information (Kantardzic, 2003). Some data mining applications are focused on outlier detection and they are the essential result of a data-analysis (Sane & Ghatol, 2006).

The outlier detection techniques find applications in credit card fraud, network robustness analysis, network intrusion detection, financial applications and marketing (Han & Kamber, 2001). A more exhaustive list of applications that exploit outlier detection is provided below (Hodge, 2004):

- Fraud detection: fraudulent applications for credit cards, state benefits or fraudulent usage of credit cards or mobile phones.
- Loan application processing: fraudulent applications or potentially problematical customers.
- Intrusion detection, such as unauthorized access in computer networks.

Source: Advances in Robotics, Automation and Control, Book edited by: Jesús Arámburo and Antonio Ramírez Treviño,
ISBN 78-953-7619-16-9, pp. 472, October 2008, I-Tech, Vienna, Austria

- Activity monitoring: for instance the detection of mobile phone fraud by monitoring phone activity or suspicious trades in the equity markets.
- Network performance: monitoring of the performance of computer networks, for example to detect network bottlenecks.
- Fault diagnosis: processes monitoring to detect faults for instance in motors, generators, pipelines.
- Structural defect detection, such as monitoring of manufacturing lines to detect faulty production runs.
- Satellite image analysis: identification of novel features or misclassified features.
- Detecting novelties in images (for robot neotaxis or surveillance systems).
- Motion segmentation: such as detection of the features of moving images independently on the background.
- Time-series monitoring: monitoring of safety critical applications such as drilling or high-speed milling.
- Medical condition monitoring (such as heart rate monitors).
- Pharmaceutical research (identifying novel molecular structures).
- Detecting novelty in text. To detect the onset of news stories, for topic detection and tracking or for traders to pinpoint equity, commodities.
- Detecting unexpected entries in databases (in data mining application, to the aim of detecting errors, frauds or valid but unexpected entries).
- Detecting mislabeled data in a training data set.

How the outlier detection system deals with the outlier depends on the application area.
A system should use a classification algorithm that is robust to outliers to model data with naturally occurring outlier points. In any case the system must detect outlier in real time and alert the system administrator. Once the situation has been handled, the anomalous reading may be separately stored for comparison with any new case but would probably not be stored with the main system data as these techniques tend to model normality and use outliers to detect anomalies (Hodge, 2004).

2. Traditional approaches

The salient traditional approaches to outlier detection can be classified as either distribution-based, depth-based, clustering, distance-based or density-based.

2.1 Distribution-based method

These methods are typically found in statistics textbooks. They deploy some standard distribution model (Normal, Poisson, etc.) and flag as outliers those data which deviate from the model. However, most distribution models typically apply directly to the future space and are univariate i.e. have very few degrees of freedom. Thus, they are unsuitable even for moderately high-dimensional data sets. Furthermore, for arbitrary data sets without any prior knowledge of the distribution of points, expensive tests are required to determine which model best fits the data, if any! (Papadimitriou et al., 2002) Fitting the data with standard distributions is costly and may not produce satisfactory results. The most popular distribution is the Gaussian function. (Kim & Cho, 2006).
A method was proposed by Grubbs which calculates a Z value as the difference between the mean value for the attribute and the query value divided by the standard deviation for the

attribute, where the mean and the standard deviation are calculated from all attribute values including the query value. The Z value for the query is compared with a 1% or 5% significance level. The technique requires no pre-defined parameters as all parameters are directly derived from the data. However, the success of this approach heavily depends on the number of exemplars in the data set. The higher the number of records, the more statistically representative the sample is likely to be (Grubbs, 1969).
A Gaussian mixture model (GMM) and computed outlierness was proposed based on how much a data point deviates from the model (Roberts & Tarassenko, 1995).
The GMM is represented by equation (1):

$$P(t|x)=\sum_{j=1}^{M}\alpha_j(x)j_j(t|x) \tag{1}$$

where M is the number of kernels (ϕ), $\alpha_j(x)$ the mixing coefficients, **x** the input vector and **t** the target vector.
A Gaussian probability density function is defined by equation (2):

$$\phi_j(t \mid x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma_j^d(x)} e^{\left\{-\frac{\|t-\mu_j(x)\|^2}{2\sigma_j^2(x)}\right\}} \tag{2}$$

where d is the dimensionality of input space, σ is the smoothing parameter, $\mu_j(x)$ represents the centre of jth kernel and $\sigma_j^2(x)$ is the variance.

Another model was proposed by Laurikkala et al.(Laurikkala et al., 2000) in probably one of the simplest statistical outlier detection techniques. This method uses informal box plots to pinpoint outliers in both univariate and multivariate datasets, produces a graphical representation and allows a human auditor to visually pinpoint the outlying points. This approach can handle real-valued, ordinal and categorical attributes.
Distribution based methods have some advantages such as mathematical justification and fast evaluation once they are built. However, they also have important drawbacks, such as the need for assuming a distribution and a considerable complexity for high dimensional problems.
Indeed statistical models are generally suited to quantitative real-valued data sets at the very least quantitative ordinal data distributions, where the ordinal data can be transformed into suitable numerical values through statistical processing. This fact limits their applicability and increases the processing time if complex data transformations are necessary before processing (Hodge, 2004).

2.2 Depth-based method

This approach is based on computational geometry and computes different layers of k-d convex hulls (Johnson et al., 1998). Based on some definition of depth, data objects are organized in convex hull layers in data space according to peeling depth and outliers are expected to be found from data objects with shallow depth values. In theory, depth-based methods could work in high dimensional data space. However, due to relying on the computational of k-d convex hulls, these techniques have a lower bound complexity of

$\Omega(N^{k/2})$, where N is number of data objects and k is the dimensionality of the dataset. This makes these techniques infeasible for large dataset with high dimensionality (He et al., 2002).

2.3 Clustering

Clustering is a basic method to detect potential outliers. From the viewpoint of a clustering algorithm, potential outliers are data which are not located in any cluster. Furthermore, if a cluster significantly differs from other clusters, the objects in this cluster might be outliers. A clustering algorithm should satisfy three important requirements (Birant & Kut, 2006):

- Discovery of clusters with arbitrary shape;
- Good efficiency on large databases
- Some heuristics to determine the input parameters.

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \| x_i - c_j \|^2 \tag{3}$$

where C is the total number of clusters, N is the total number of data, m is any real number greater than 1, u_{ij} is the degree of membership of x_i to the j-th cluster, x_i is the i-th of the d-dimensional measured data, c_j is the d-dimensional center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the above cited objective function, with the update of membership degree u_{ij} and the cluster centers c_j respectively given by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|}\right)^{\frac{2}{m-1}}} \tag{4}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{5}$$

The iteration stops when the following condition holds:

$$\max_{ij} \left\{ | u_{ij}^{(k+1)} - u_{ij}^{(k)} | \right\} < \varepsilon \tag{6}$$

where ε is a termination criterion lying in the range [0, 1] whereas k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m.

A quite popular technique used for outlier detection is the introduction of a new variable called *credibility* of a vector. This method is proposed by Chintalapudi and Kam (Chintalapudi & Kam, 1998) with regard to the above described fuzzy clustering algorithms. Credibility measures the "how typical" the vector is with respect to the entire data set. An

outlier is expected to have a low value of credibility compared to a non-outlier. The use of the new variable leads to the Credibilistic Fuzzy C Means algorithm. Two formulations are made for the credibility variable, based on the mean nearest statistical distance. Simulations demonstrate that the proposed schemes are robust with respect to the presence of outliers and with respect to the parameter values needed to implement the algorithm.

Most clustering algorithms, especially those developed in the context of Knowledge Discovery in Databases (KDD) (e.g. CLARANS (Ng&Han, 1994), DBSCAN (Ester et al., 1998), BIRCH (Zhang et al., 1996), STING (Wang et al., 1997), Wave Cluster (Sheikholeslami et al., 1998), DenClue (Hinneburg&Keim, 1998), CLIQUE (Aggarwal et al., 1998) , OPTICS (Ankerst et al.,1999), PROCLUS (Aggarwal et al., 1999)) are to some extent capable of handling exceptions.

However, since the main objective of a clustering algorithm is to find clusters, they are developed to optimize the outlier detection. The exceptions (called "noise" in the context of clustering) are typically just tolerated or ignored when producing the clustering result. Even if the outliers are not ignored, the notions of outliers are essentially binary and there are no qualification as to how outlying an object is (Breunig et al., 2000).

2.4 Distance-based method

The notion of distance-based (DB) outlier is been defined by Knorr and Ng (1988):

An object O in a dataset T is a DB(p,D)-outlier if at least fraction p of the objects in T lie greater than distance D from O.

The concept of DB-outlier is well defined for any dimensional dataset. The parameter *p* is the minimum fraction of objects in a data space that must be outside an outlier D-neighborhood (Li & Hiroyuki, 2007).

This notion generalizes many concepts from distribution-based approach and better faces computational complexity. It is further extended based on the distance of a point from its k-th nearest neighbor (Ramasmamy et al., 2000).

After ranking points by the distance to its *k*-th nearest neighbor, the top *k* points are identified as outliers. Alternatively, in the algorithm proposed by Angiulli and Pizzuti (Angiulli & Pizzuti,2000), the outlier factor of each data point is computed as the sum of distances from its *k* nearest neighbors.

Both the method proposed in (Matsumoto et al., 2007) and the Mahalanobis outlier analysis (MOA) (Marquez et al., 2002) are distance-based approaches which exploit Mahalanobis distance as outlying degree of each data point.

In 1936 P.C. Mahalanobis introduced a distance measure (Mahalanobis, 1936) which is based on correlations between variables by which different patterns can be identified and analyzed and provides a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements.

Formally, the Mahalanobis distance from a group of values with mean $\mu=(\mu_1, \mu_2, \mu_3, \ldots, \mu_p)^T$ and covariance matrix Σ for a multivariate vector $x=(x_1, x_2, x_3, \ldots, x_p)^T$ is defined as:

$$D_M = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{7}$$

Mahalanobis distance can also be defined as dissimilarity measure between two random vectors **x** and **y** of the same distribution with the covariance matrix P:

$$d(x,y) = \sqrt{(x-y)^T P^{-1}(x-y)} \tag{8}$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. Finally if the covariance matrix is diagonal, then the resulting distance measure is called the normalized Euclidean distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{P} \frac{(x_i - y_i)^2}{\sigma_i^2}} \tag{9}$$

where σ_i is the standard deviation of the x_i over the sample set.

Mahalanobis distance is computed on the basis of the variance of data points. It describes the distance between each data point and the center of mass.

When one data point is on the center of mass, its Mahalanobis distance is zero, and when one data point is distant from the center of mass, its Mahalanobis distance is more than zero. Therefore, datapoints that are located far away from the center of mass are considered outliers (Matsumoto et al., 2007).

2.5 Density-based method

This method assigns to each object a degree to be an outlier. This degree is called the local outlier factor (LOF) of an object. It is "local" in the sense that the degree depends on how isolated the object is with respect to the surrounding neighborhood.

In LOF algorithm, outliers are data objects with high LOF values whereas data objects with low LOF values are likely to be normal with respect to their neighborhood. High LOF is an indication of low-density neighborhood and hence high potential of being outlier (Mansur & Noor, 2005). In order to understand LOF method, it is necessary to define some auxiliary notions. (Breunig et al., 2000).

Definition 1: (k-distance of an object p).

For any positive integer k, the k-distance of object p, denoted as k-distance (p), is defined as the distance d(p,o) between p and an object o∈D such that:

i. For at least k objects o′∈ D\ {p} it holds that d(p,o′)≤d(p,o) and
ii. For at most k-1 objects o′∈ D\ {p} it holds that d(p,o′)<d(p,o).

Definition 2: (k-distance neighborhood of an object p).

Given the k-distance of p, the k-distance neighborhood of p contains every object whose distance from p is not greater than the k-distance, i.e.

$$N_{k-dis\tan ce(p)}(p) = N_k(p) = \{q \in D \setminus \{p\} \mid d(p,q) \le k - dis\tan ce(p)\} \tag{10}$$

These objects q are called the k-nearest neighbors of p.

Definition 3: (reachability distance of an object p w.r.t. object o).

Let k be a natural number. The reachability distance of object p with respect to object o is defined as

$$reach - dist_k = \max\{k - dis\tan ce(o), d(p,o)\} \tag{11}$$

The higher the value of k, the more similar the reachability distances for objects within the same neighborhood.

In a typical density-based clustering algorithm, there are two parameters that define the notion of density: a parameter *MinPts* specifying a minimum number of objects and a parameter specifying a volume. These two parameters determine a density threshold for the clustering algorithms to operate. That is, objects or regions are connected if their neighborhood densities exceed the given density threshold. To detect density-based outliers, however, it is necessary to compare the densities of different sets of objects, which means that the density of sets of objects must be dynamically determined. An idea is to keep MinPts as the only parameter and use the values reach-$dist_{MinPts}$(p,o), for o $\in N_{MinPts}$(p).

Definition 4: (local reachability density of an object p).

The local reachability density of p is defined as:

$$lrd_{MinPts}(p) = \frac{1}{\frac{\sum_{o \in N_{MinPts}(p)} reach - dist_{MinPts}(p,o)}{\left| N_{MinPts}(p) \right|}} \tag{12}$$

Intuitively, the local reachability density of an object p is the inverse of the average reachability distance based on the MinPts-nearest neighbors of p.

Definition 5: (local outlier factor of an object p)

The local outlier factor of p is defined as:

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{| N_{MinPts}(p) |} \tag{13}$$

The outlier factor of object p captures the degree to which we call p an outlier. It is the average of the ratio of the local reachability density of p and those of p's MinPts-nearest neighbors.

It is possible to know the range where LOF lies. By defining ε as

$$\varepsilon = \frac{\max\{reach - dist(p,q)\}}{\min\{reach - dist(p,q)\} - 1} \tag{14}$$

it can be demonstrated that (Breunig et al., 2000):

$$\frac{1}{1+\varepsilon} \le LOF(p) \le 1 + \varepsilon \tag{15}$$

3. Wavelets-based outlier detection approach

The *wavelet transform* or *wavelet analysis* is probably one of the most recent solutions to overcome the shortcomings of the Fourier transform. In wavelet analysis the use of a fully scalable modulated window solves the signal-cutting problem. The window is shifted along the signal and for every position the spectrum is calculated. This process is repeated many times with a slightly shorter (or longer) window for every new cycle. The processing result

is a collection of time-frequency representations of the signal corresponding to different resolutions.

The wavelet transform is an operation that transforms a function by integrating it with modified versions of some kernel function. (Combes et al., 1989). The kernel function is called the *mother wavelet*, and the modifications are translations and compressions of the mother wavelet. A function *g(t)* can be a mother wavelet if it is *admissible*, i.e. if the following condition holds (Grossmannn & Morlet, 1984):

$$c_g \equiv \int_{-\infty}^{+\infty} \frac{|G(\omega)|^2}{|\omega|} d\omega < \infty \tag{16}$$

where $G(\omega)$ is the Fourier transform of $g(t)$. The constant c_g represent the admissibility of the function that must be finite for inversion of the wavelet transform. Any admissible function can be a mother wavelet (Weiss, 1994).

Wavelet transforms are classified into discrete wavelet transforms and continuous wavelet transforms on the basis of the applications and data.

Outlier detection by means of the wavelet transform is a recent study area. (Bruce et al., 1994) (Wang, 1995) (Dorst, 1999) and can be an alternative to the previously discussed methods because the data do not need to belong to a known distribution and the discontinuities in the processed data, that can often correspond to outliers, are easily detected. (Kern et al., 2005).

For instance Yu et al. (Yu et al., 2002) introduced a new outlier detection approach, called FindOut (which stands for *Find Outliers)*, based on wavelet transform. FindOut identifies outliers by removing clusters from the original data. The main idea is to apply signal processing techniques to transform the space and find the dense regions in the transformed space. The remaining objects in the non-dense regions are labeled as outliers. The primary motivation for applying signal-processing techniques to spatial datasets comes from the observation that the multidimensional data points can be represented as a d-dimensional signal. Wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-bands; this method use two filters: a high-pass filter and a low-pass filter. One important step of wavelet transform is to convolve its input with a low-pass filter that has the main property of removing noise (outlier). The low-pass filter removes the outliers and smoothes the input data. However in FindOut the objective is to find outliers and not remove them so the idea is to remove the clusters from the original data and thus identify the outliers.

4. Outlier detection using artificial intelligence techniques

When dealing with industrial automation, where data coming from the production field are collected with different and heterogeneous means, the occurrence of outliers is more the rule than the exception. Standard outlier detection methods fail to detect outliers in industrial data because of the high dimensionality of the data. In these cases, the use of artificial intelligence techniques has received increasing attention in the scientific and industrial community, as the application of these techniques shows the advantage of requiring poor or no *a priori* theoretical assumption on the considered data. Moreover their implementation is relatively simple and with no apparent limitation on the dimensionality of the data.

4.1 Neural networks

In 1943, McCulloch and Pitts introduced the idea of an artificial neuron to process data. In the 50ies this work was advanced by arranging neurons in layers. Although learning rules to cope with multiple layers of perceptrons were not developed until later, this work formed the basis of the Multi-Layer Perceptron (MLP) that is used today.

Another kind of neural network that is frequently used is the Radial Basis Function (RBF), which exploits gaussian activation functions in the first (or sometimes called hidden) layer.

Once the inputs and the outputs have been defined, it is useful to see if the data set contains any points that violate this limits. If there are many similar examples for a given input pattern, an outlier can be classified as the one which is furthest from the median value.

Other methods that can be applied to detect outliers are the Principle Component Analysis (PCA) and Partial Least Squares (PLS). Outliers can be found by investigating points at the edges of the previously created clusters. (Tenner et al., 1999).

Liu and Gader indicated that including outlier samples in training data and using more hidden nodes than required for classification for MLP and BRF networks and proceeding an RBF with principal Component decomposition can achieve outlier rejection. The further addition of a regularization term to the PCA-RBF can achieve an outlier rejection performance equivalent or better than that of other networks without training on outliers (Liu & Gader, 2000).

Williams et al. propose to apply the so-called Replicator Neural Network (RNN) for outlier detection. RNN are multi-layer perceptron neural networks with three hidden layers and the same number of output neurons and input neurons to model the data. The input variables are also the output variables so that the RNN forms compressed model of data during training. A measure of outlyingness of individuals is developed as the reconstruction error of individual data points. (Hawkins et al., 2002). This method is often compared with other methods, in particle with Hadi94 (Hadi, 1994) and Donoho-Stahel (Knorr et al., 2001).

The Hadi94 is a parametric bulk outlier detection method for multivariate data and Donoho-Stahel uses the outlyingness measure compute by the Donoho-Stahel (Rousseeuw & Leroy, 1997) estimator of location and scatter. These two methods perform well on large and complex data sets. However these are parametric methods and lack of the flexibility of non-parametric methods, such as Minimum Message Length (MLL) (Oliver et al., 1996) and RNN. MLL clustering works well for scattered outlier while RNN degrades with datasets containing radial outliers. However RNN performs satisfactory for small and large datasets (Neural Network methods often have difficulty with such smaller datasets) (Williams et al., 2002).

The self-organizing map (SOM) is an artificial neural networks, which is trained by using unsupervised learning in order to produce a low dimensional representation of the training samples while preserving the topological properties of the input space. (Munoz & Muruzabal, 1997).

Nag et al. (Nag et al., 2005) proposed a SOM based method for outlier detection, which identifies the multidimensional outliers and provides information about the entire outlier neighborhood. The SOM based outlier detection method is non-parametric and can be used to detect outliers from large multidimensional datasets. This method has the advantages that does not require any a priori assumption on the variable, is easy to implement and does not have problems with dimensionality of data.

4.2 Support Vector Machine (SVM)

Support Vector Machine, introduced by V. Vapnik (Vapnik, 1982), is a method for creating functions from a set of labelled training data. The function can be a classification function or a general regression function. In classification tasks, SVMs operate by finding a hypersurface in the space of the possible inputs, which separates the samples belonging to different classes. Such division is chosen so as to have the largest distance from the hypersurface to the nearest of the positive and negative examples (Vapnik, 1998).

Jordaan and Smits (Jordaan & Smits, 2004) propose a robust model-based outlier detection approach that exploits the characteristics of the support vectors extracted by the SVM method (Cortes & Vapnik, 1995). This method makes use of several models of varying complexity to detect outliers based on the characteristics of the support vectors obtained from SVM-models. This approach has the advantage that the decision does not depend on the quality of a single model, which adds to the robustness of the approach. Furthermore, since it is an iterative approach, the most severe outliers are firstly removed. This allows the models in the next iteration to learn from "cleaner" data and thus reveal outliers that were masked in the initial model. The need for several iterations as well as the use of models, however, makes the on-line application of this method difficult for the not negligible computational burden. Moreover, if the data to be on-line processed come from dynamic systems, which tend to change (more or less rapidly) their conditions through time, the models update is required and this can furtherly slow the outlier detection.

4.3 Fuzzy logic

Fuzzy Logic (FL) is linked with the theory of fuzzy sets, a theory which relates to classes of objects with un-sharp boundaries in which membership is a matter of degree.

Fuzzy theory is essential and is applicable to many systems — from consumer products like washing machines or refrigerators to big systems like trains or subways. Recently, fuzzy theory has been a strong tool for combining new theories (called soft computing) such as genetic algorithms or neural networks to get knowledge from real data (Melin & Castillo, 2008).

Fuzzy logic is conceptually easy to understand, tolerant of imprecise data and flexible. Moreover this method can model non-linear functions of arbitrary complexity and it is based on natural language. Natural language has been shaped by thousands of years of human history to be convenient and efficient. Since fuzzy logic is built atop the structures of qualitative description used in everyday language, fuzzy logic is easy to use (Baldwin, 1978).

Fuzzy inference system (FIS) (Ross, 2004) is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or pattern discerned. The process of fuzzy inference involves: membership functions (MF), a curve that defines how each point in the input space is mapped to a membership value or degree of membership between 0 and 1; fuzzy logic operators (and, or, not); if-then rules. Since decisions are based on the testing of all of the rules in an FIS, the rules must be combined in some manner in order to make decision. Aggregation is the process by which the fuzzy sets that represents the outputs of each rule are combined into a single fuzzy set. Aggregation only occurs once for each output variable, just prior to the final step, defuzzification (Patyra & Mlynek, 1996).

Due to the linguistic formulation of its rule basis, the FIS provides an optimal tool to combine more criteria among those that were above illustrated according to a reasoning that

is very similar to the human one. So doing, in practical application, the knowledge of the technical expert personnel can easily be exploited by the system designer.
In the following, an exemplar method will be proposed which exploits a fuzzy inference system (FIS) in order to combine several outliers detection strategies by jointly evaluating four features of a particular datum within a series of measurements (Cateni et al., 2007).
For each pattern the following four features are extracted and fed as input of the FIS:

- Distance between each element and the centroid of the overall distribution normalized with respect to the average value. (dist)
- Fraction of the total number of elements that are near to the pattern itself. (n-points)
- Mean distance between the considered pattern and the remaining patterns normalized with respect to the maximum value. (memb-deg)
- Degree of membership of the patterns to the cluster to which it has been assigned by the preliminary fuzzy c-means clustering stage. (mean-dist)

In geometry the *centroid* of an object X in n-dimensional space is the intersection of all hyperplanes that divide X into two parts of equal moment about the hyperplane. Roughly speaking, the centroid is a sort of "average" of all points of X.
The FIS is of Mandami type (Mandami & Assilian, 1975) and the FIS output variable, named *outlier-index* (outindx), is defined in the range [0;1]. The output function provides an indication on the risk that the considered pattern is an outlier.
The inference rules relating the output variable to the four inputs is formulated through a set of 6 fuzzy rules, that are listed below:

1. IF (dist is very high) AND (n-points is very small) AND (memb-deg is low) AND (mean–dist is big) THEN (outindx is very high).
2. IF (dist is medium) AND (n-points is small) AND (memb-deg is quite low) AND (mean–dist is small) THEN (outindx is quite high).
3. IF (dist is low) AND (n-points is medium) AND (memb-deg is quite low) AND (mean–dist is very small) THEN (outindx is low).
4. IF (dist is medium) AND (n-points is very small) AND (memb-deg is quite low) AND (mean–dist is small) THEN (outindx is quite high).
5. IF (dist is low) AND (n-points is small) AND (memb-deg is high) AND (mean–dist is quite big) THEN (outindx is low).
6. IF (dist is low) AND (n-points is medium) AND (memb-deg is high) AND (mean–dist is small) THEN (outindx is low).

Figure 1 depicts a scheme of the proposed method. An outlier is detected when its outlier index overcome a prefixed threshold.

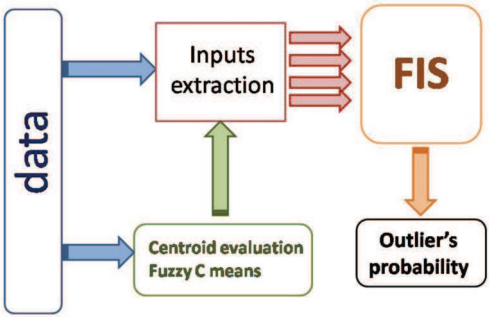

Fig. 1. Block diagram depicting the overall system for outliers detection.

5. Results

The proposed outlier detection method, which does not require a priori assumption on the data, has been tested in the preprocessing of data provided by a steelmaking industry, where outliers can provide indications on malfunctionings or anomalous process conditions. The method has been tested by considering that the technical personnel provided indications of those values that should be considered as outliers.

In this work two applications are proposed that use two different variables that are important to determine the quality of steel and the final destination. The variables are composed by 100 samples normalized respect their mean value.

The performance of the fuzzy logic-based method has been compared with Grubbs test and Local Outlier Factor (LOF) techniques, that are considered among the most important and widely adopted traditional outlier detection methods. The results show that the fuzzy logic-based method outperforms the other approaches, but, on the other hand, the required computational time is approximately ten times greater than the time required by traditional methods, due to the increased complexity of the FIS-based evaluation.

In particular, in the first exemplar application, the considered variable represents the concentration of a chemical element extracted from the analysis made on molten steel. In this piece of database there are five outliers. The samples that are considered outliers are 3, 8, 16, 35 and 92 referring respectively to following values:

0.9349 0.9385 1.0455 0.9626 0.9541.

Figure 2 shows the result of Grubbs test. It clearly appears that only two anomalous samples have been recognized as outliers (the 3rd and 8th samples).

Figure 3 shows the result of LOF technique: only three samples are exactly classified as outliers but there are two samples that this method does not recognize as outliers.

Finally, in figure 4, is shown the result using fuzzy logic method. It clearly appears that all outliers have been correctly recognized.

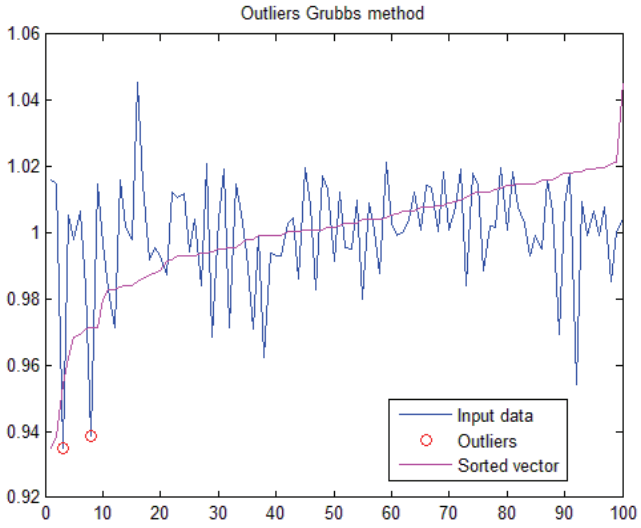

Fig. 2. First example using Grubbs method.

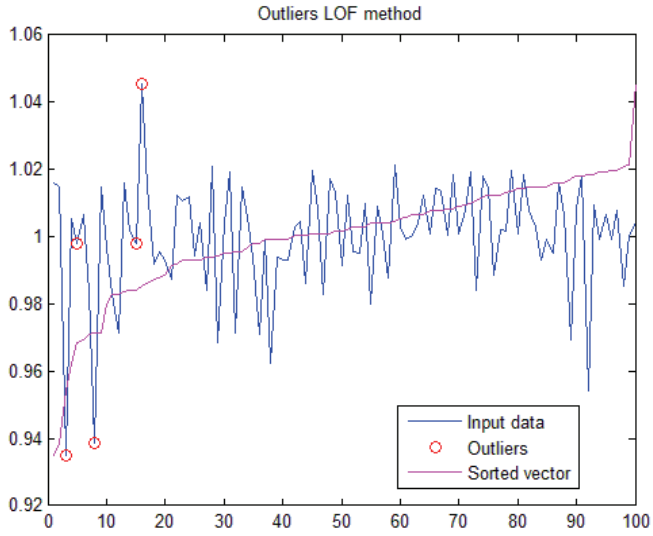

Fig. 3. First example using Local Outlier Factor method.

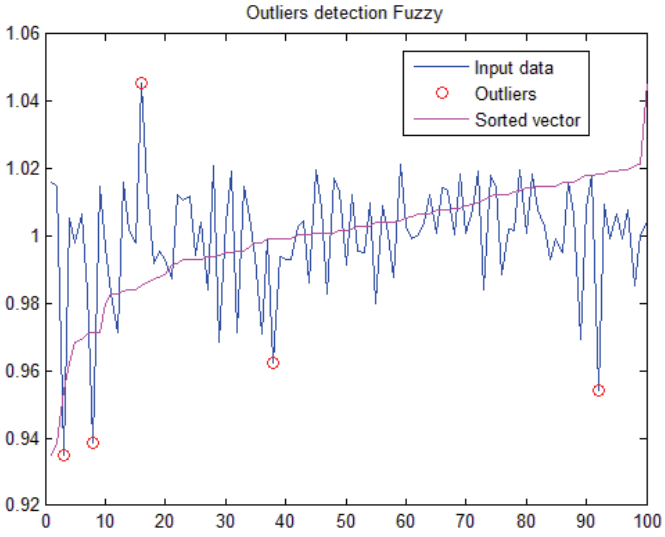

Fig. 4. First example using Fuzzy Logic method.

Similar results are obtained in the second application.
The second examined variable is referred to a sensor detection made within the process. In this case, the samples considered outliers are 3,5,7,40 and 60 referring respectively to following values:

0.4730 0.4779 0.4779 0.4680 1.9217.

The Grubbs test does not detect any outliers. The result is shown clearly in figure 5.

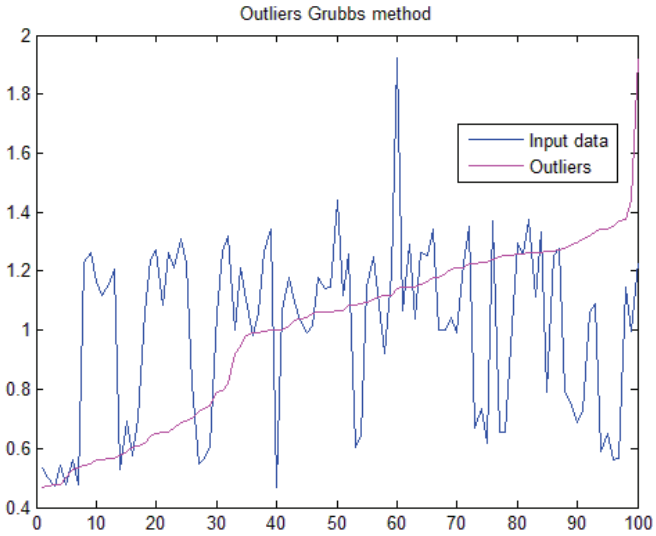

Fig. 5. Second example using Grubbs method.

The Local Outlier Factor technique detects four outliers but only one is really an outlier. The result is shown in figure 6.

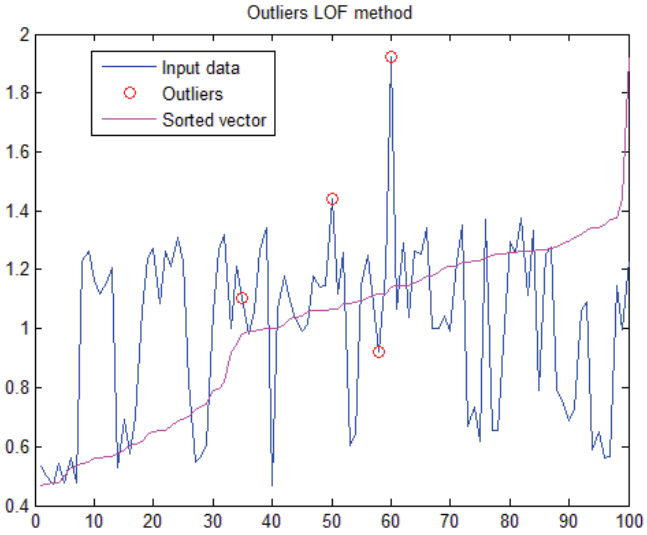

Fig. 6. Second example using Local Outlier Factor method.

The proposed method, as show figure 7, recognize all the outliers which are present in data.

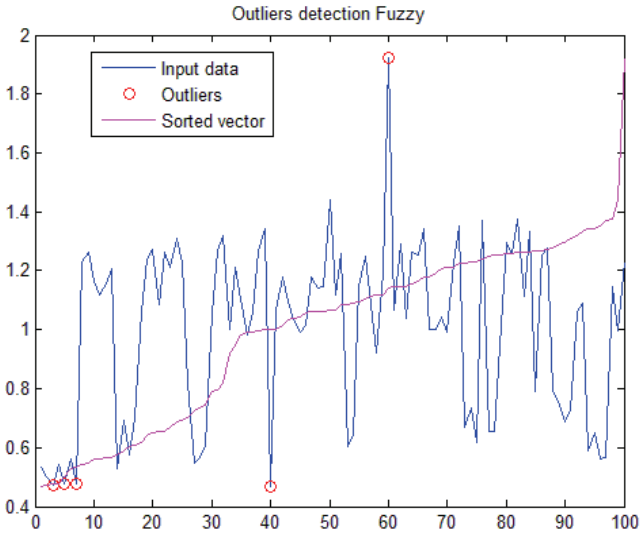

Fig. 7. Second example using Fuzzy Logic method.

6. Conclusions and future work

A description of traditional approaches and of the most widely used methods within each category has been provided. As standard outlier detection methods fail to detect outliers in industrial data, the use of artificial intelligence techniques has also been proposed, because it presents the advantage of requiring poor or no a priori assumption on the considered data.
A procedure for outlier detection in a database has been proposed which exploits a Fuzzy Inference System in order to evaluate four features for a pattern that characterize its location within the database. The system has been tested on a real industrial application, where outliers can provide indications on malfunctionings or anomalous process conditions. The presented results clearly demonstrate that the Fuzzy Logic-based method outperforms the most widely adopted the traditional methods.
Future work on the FIS-based outliers detection strategy will concern the algorithm optimization in order to improve its efficiency and its on-line implementation. Moreover further tests will be performed on different applications.

7. References

Aggarwal, C.C.; Procopiuc, C.; Wolf, J.L.; Yu, P.S. & Park, J.S. (1999). Fast algorithms for projected clustering, *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 61–72, Philadephia, Pennsylvania, U.S.A.

Aggarwal, R.; Gehrke, J.; Gunopulos, D. & Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, *Proceedings of*

ACM SIGMOD International Conference on Management of Data, pp. 94-105, Seattle, WA.

Ankerst,M.; Breunig, M.M.; Kriegel, H.P. & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure, *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pp. 49–60, June 1999, Philadelphia, Pennsylvania, U.S.A.

Baldwin, J.F. (1978). Fuzzy Logic and Fuzzy Reasoning. *International Journal of Man-Machine Studies*, Vol. 11, pp. 465-480.

Barnet, V. & Lewis, T. (1994), *Outliers in statistical data,* John Wiley, ISBN 0-471-93094-6, Chichester.

Birant, D.& Kut, A. (2006). Spatio-Temporal Detection in Large Databases, *Proceedings of the 28th International Conference Information Technology Interfaces ITI 2006* June 19-22, Croatia.

Breunig, M.M; Kriegel, H.P.; Ng, R.T. & Sander, J. (2000) LOF: Identifying density-based local outliers, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93-104, May 2000, Dallas.

Bruce, A.G.; Donoho L.G.; Gao, H.Y. & Martin R.D. (2004). Denoising and robust nonlinear wavelet analysis, *SPIE Proceedings Wavelet Applications,* Vol. 2242, pp. 335-336, Harald H.San (ed), The International Society for Optical Engineering (SPIE), Orlando, FL.

Cateni, S.; Colla, V. & Vannucci, M. (2007). A fuzzy logic-based method for outlier detection, *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*, pp.561-566, Innsbruck, Austria.

Chen, Z.; Fu, A. & Tang, J., (2002). *Detection of outliered Patterns*, Dept. of CSE, Chinese University of Hong Kong.

Chintalapudi, K. & Kam,M. (1998), The credibilistic fuzzy c means clustering algorithm, *IEEE Fuzzy Systems Proceedings*, Vol. 2, pp. 2034-2039.

Combes, J.M.; Grossman A. & Tchamitchian P. (1989) *Wavelets: Time-Frequency Methods and Phase Space,* Second Edition, Springer-Verlag, New York.

Cortes, C. & Vapnik, V. (1995). Support vector networks, *Machine Learning*, Vol. 20, pp. 273–297.

Dorst, L. (1999) Current and addional procedures for superconducting gravimer data at the main tidal frequency. *Graduation Report,* Delft University of Technology, Delft.

Ester, M.; Kriegel, H.P., Sander, J. & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proceedings of the 2nd conference on Proceedings of the 25th IASTED International Conference on Knowledge Discovery and Data Mining*, pp. 226-231, Portland.

Grossmann, A. & Morlet, J. (1984) *Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape,* SIAM J.Math. Anal. Vol. 15, N° 4, pp. 723-736.

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

Hadi, A. (1994) A modification of a method for the detection of outliers in multivariate samples, *Journal of Royal Statistical Society B,* Vol. 56, N°2.

Han, J. & Kamber M. (2001) *Data Minings Concepts and Techniques,* Morgan Kauffman Publisdhers.

Hawkins, D. (1980), *Identification of Outliers,* Chapman and Hall, London.

Hawkins, S.; He, X.; Williams, G.J. & Baxter, R.A. (2002). Outlier detection using replicator neural networks. *Proceedings of the 5th international conference on Knowledge Discovery and Data Warehousing*.

He, Z.; Xu, X. & Deng, S. (2002). Discovering cluster-based local outliers. *Pattern Recognition Letters*, Vol. 24, N. 9-10, pp. 1641-1650, ISSN 0167-8655.

Hinneburg, A. & Keim D. A. (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise, *Proceedings of the 4th international conference on Knowledge Discovery and Data Mining*, pp. 58-65, New York City, NY.

Hodge, V.J. (2004), *A survey of outlier detection methodologies*, Kluver Academic Publishers, Netherlands, January 2004.

Jiemenez-Marquez, S.A.; Lacroix, L. & Thibault, J. (2002) *Statistical data validation methods for large cheese plant database*. J.Dayry Sci., Vol.85, N°9, pp.2081-2097, Sep 2002.

Johnson, T.; Knok, I.; Ng, R. (1998). Fast computation of 2-dimensional depth contours, *Proceedings of 4thInternational Conference on Knowledge Discovery &Data Mining,* pp.224-228, New York, August 1998.

Jordaan, E.M. & Smits, G.F.(2004) Robust Outlier Detection using SVM Regression, *Proceeding. of the IEEE, International Joint Conference on Neural Networks*, Vol.3, pp. 2017-2022, July 2004.

Kantardzic, M. (2003). Data mining Concepts, Models, Methods and Algorithms. *IEEE Transactions on neural networks,* Vol.14, N. 2, March 2003.

Kern, M.; Preimesberger, T; Allesch, M.; Pail, R.; Bouman, J. & Koop, R. (2005). Outlier detection algorithms and their performance in GOCE gravity field processing, *Journal of Geodesy,* Vol. 78, pp. 509-519, January 2005.

Kim, S. & Cho, S. (2006). Prototype based outlier detection. *Proceedings of International JoinConference on Neural Networks,* ISBN 0-7803-9490-9, Vancouver, BC, Canada, 16-21 July 2006.

Knorr, E.M..; Ng, R. (1988). Algorithms for Mining Distance-Based Outliers in Large Datasets., *Proceedings of VLDB,* pp.392-403.

Knorr, E.M.; Ng, R.T. & Zamar, R.H. (2001) Robust Space Transformation for Distance-based Operations, *Proceeding of the 7th International Conference on Knowledge Discovery and Data Mining KD001,* pp. 126-135.

Laurikkala, J.; Juhola, M. & Kentala, E. (2000) Informal Identification of Outliers in Medical Data. *Proceeding in the Fifth International Workshop on Intelligent Data Analysis n Medicine and Pharmacology IDAMAP-2000 Berlin,* 22 August. *Organized as a workshop of the 14th European Conference on Artificial Intelligence* ECAI-2000.

Li, Y. & Hiroyuki, K. (2007). Example-Based DB-Outlier Detection from high Dimensional Datasets. *Proceedings of DEWS*.

Liu, J. & Gader, P. (2000). Outlier rejection with MLPs and variants of RBF Networks, *Proceedings of the 15th IEEE International Conference on Pattern Recognition*, Vol.2, pp. 680-683, 3-7 September, 2000, Missouri.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India,* pp. 49-55.

Mamdani, E.H. & Assilian, S. (1975) An experiment in linguistic synthesis with a fuzzy logic controller, *International Journal of Man-Machine Studies*, Vol. 7, No. 1, pp. 1-13.

Mansur M.O. & Mohd. Noor Md. Sap (2005), Outlier detection technique in data mining : a research perspective, *Proceedings of the postgraduate annual research seminar*.

Matsumoto, S.; Kamei, Y; Monden, A. & Matsumoto K. (2007) Comparison of Outlier Detection Methods in Fault-proneness Models. *Proceedings of the 1st International Symposium on Empirical Software Engineering and Measurement* (ESEM2007), pp.461-463, September 2007.

Melin, P. & Castillo, O. (2008). *Fuzzy logic: theory and applications*, Springer.

Moore, D.S. & McCabe G.P. (1999), *Introduction to the Practice of Statistics.* , Freeman & Company.

Munoz, A. & Muruzabal, J. (1997) *Self organizing maps for outliers detection*, Elsevier, Neurocomputing, N°18, pp.33-60, August 1997.

Nag, A.K.; Mitra, A. & Mitra, S.(2005), Multiple outlier Detection in Multivariate Data Using Self-Organizing Maps Title, Computational Statistical, N.20, pp.245-264.

Ng R. T. & Han J. (1994) Efficient and Effective Clustering Methods for Spatial Data Mining, *Proceeding of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, Morgan Kaufmann Publishers, pp. 144-155, San Francisco, CA.

Oliver, J.J.; Baxter, A. & Wallace, C.S. (1996), Unsupervised Learning using MML, *Proceedings of the 13th International Conference (ICML),* pp. 364-372, Morgan Kaufmann Publishers, San Francisco, CA.

Papadimitriou, S.; Kitawaga, H.; Gibbons, P.B. & Faloutsos C. (2002), LOCI: Fast Outlier Detection Using the Local Correlation Integral, *Proceedings of the 19th International Conference on Data Engineering (ICDE),* pp. 315-326.

Patyra, M.J. & Mlynek D.J. (1996). *Fuzzy logic: Implementation and applications.*Wiley & Teubner, ISBN 047195099.

Ramasmawy R.; Rastogi R. & Kyuseok S. (2000). Efficient algorithms for mining outliers from large data sets. *Proceedings of the ACM SIGMOD International Conference on Management of Data,* pp.427-438, ISBN 1-58113-217-4, Dallas, Texas, United States.

Roberts, S. & Tarassenko, L. (1995). A Probabilistic Resource Allocating Network for Novelty Detection. *Neural Computation,* Vol. 6, N. 2, , 1995, pp. 270-284.

Ross, Timothy J. (2004). *Fuzzy logic with engineering appplications*, John Wiley & sons ltd, England.

Rousseeuw, P. & Leroy, A. (1997). *Robust Regression and Outlier detection,* John Wiley & Sons.

Sane, S. & Ghatol, A. (2006), Use of Instance Tipicality for Efficient Detection of Outliers with neural network Classifiers. *Proceedings of 9thInternational Conference on Information Technology*, ISBN 0-7695-2635-7.

Sheikholeslami, G.; Chatterjee, S. & Zhang A.. (1998). WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases, *Proceedings International Conference on Very Large Data Bases*, pp. 428-439, New York, NY.

Tenner, J.; Linkens, D.A. & Bailey, T.J. (1999). Preprocessing of Industrial Process Data with Outlier detection and correction, *Proceedings of the 2nd International Conference on Intelligent Processing and Manufacturing of Materials, IEEE,* , pp. 921–926, Vol. 2, 10-15 July 1999.

Vapnik, N. (1982) *Estimation of dependence based on empirical data,* Springer-Verlag, New York.

Vapnik, N. (1998) *Statistical learning theory.* John Wiley & Sons, ISBN 0-471-03003-1, New York.

Yen, J. & Langari, R. (1998) *Intelligence control and information.* Prentice Hall, 1998.

Yu, D.; Sheikholeslami G. & Zhang, A. (2002) FindOut: Finding Outliers in Very Large Datasets. *Knowledge and Informations Systems,* vol.4, pp. 387-412, Springer-Verlag, London.

Wang, Y. (1995) *Jump and sharp cusp detection by wavelets.* Biometrika, Vol. 82, pp. 385-397.

Wang, W.; Yang J. & Muntz, R. (1997) *STING:* A Statistical Information Grid Approach to Spatial Data Mining, *Proceedings of the 23th International Conference on Very Large Data Bases*, Athens, Greece, Morgan Kaufmann Publishers, pp. 186-195, 1997, San Francisco, CA.

Weiss, L.G. (1994). Wavelets and Wideband Correlation Processing, *Signal Processing magazine* IEEE, Vol. 11, pp.13-32, ISSN 1053-5888, January 1994.

Williams, G.; Baxter, R.; He, H. & Hawkison,S. (2002). A comparative study of RNN for outlier detection in data mining, *Proceedings of the IEEE International Conference on Data Mining*, pp. 709–712, 9-12 December 2002, Australia.

Zhang, T.; Ramakrishnan, R. & Linvy M. BIRCH (1996). An Efficient Data Clustering Method for Very Large Databases, *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 103-114, ACM Press, New York.

Advances in Robotics, Automation and Control

Edited by Jesus Aramburo and Antonio Ramirez Trevino

ISBN 978-953-7619-16-9
Hard cover, 472 pages
Publisher InTech
Published online 01, October, 2008
Published in print edition October, 2008

The book presents an excellent overview of the recent developments in the different areas of Robotics, Automation and Control. Through its 24 chapters, this book presents topics related to control and robot design; it also introduces new mathematical tools and techniques devoted to improve the system modeling and control. An important point is the use of rational agents and heuristic techniques to cope with the computational complexity required for controlling complex systems. Through this book, we also find navigation and vision algorithms, automatic handwritten comprehension and speech recognition systems that will be included in the next generation of productive systems developed by man.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Silvia Cateni, Valentina Colla and Marco Vannucci (2008). Outlier Detection Methods for Industrial Applications, Advances in Robotics, Automation and Control, Jesus Aramburo and Antonio Ramirez Trevino (Ed.), ISBN: 978-953-7619-16-9, InTech, Available from: http://www.intechopen.com/books/advances_in_robotics_automation_and_control/outlier_detection_methods_for_industrial_applications

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821