International Journal of Multidisciplinary and Current Research

Research Article

ISSN: 2321-3124
Available at: http://ijmcr.com

The Detail Survey of Anomaly/Outlier Detection Methods in Data Mining

Alka P.Beldar and Vinod S.Wadne (Guide)

Department of Computer Engineering, JSPM's Imperial College of Engineering and Research, Wagholi, Pune, India

Accepted 25 May April 2015, Available online 26 May 2015, **Vol.3** (**May/June 2015 issue**)

Abstract

Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many real-world applications such as intrusion or credit card fraud detection require an effective and efficient framework to identify deviated data instances. This template provides an easier and succinct understanding and comparisons between each of the outlier detection techniques for different applications. And introduce a new efficient approach which detects outlier with imperfect labels.

Keywords: *likelihood values, uncertain data.*

1. Introduction

Data mining is a process of extracting valid, previously unknown, and ultimately comprehensible information from large datasets and using it for organizational decision making [1]. However, there a lot of problems exist in mining data in large datasets such as data redundancy, the value of attributes is not specific, data is not complete and outlier [2]. An outlier is defined as data point which is very different from the rest of the data based on some measure. Such a point often contains useful information on abnormal behavior of the system described by data [3]. Outlier detection problem is one of the very interesting problems arising recently in the data mining research. Recently, a few studies have been conducted on outlier detection for large datasets [3]. Outlier detection refers to the problem of detecting and analyzing patterns in data that does not map to expected normal behavior. These patterns are often referred to as outliers, anomalies, discordant observations, exceptions, noise, errors, novelty, damage, faults, defects, aberrations, contaminants, surprise or peculiarities in different application domains [16].

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. Outlier detection or outlier mining is the process of identifying outliers in a set of data. The outlier detection technique finds applications in credit card fraud, network robustness analysis, network intrusion detection, financial applications and marketing [3]. Thus, outlier detection and analysis is an interesting and important data mining task.

1.1 What are outliers?

Outliers are data that are outside the limits of most of your other data. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

1.2 Application scenario

General application scenarios of outlier detection are:

Supervised scenario

In some applications, training data with normal and abnormal data objects are provided .There may be multiple normal and/or abnormal classes. Often, the classification problem is highly unbalanced. This approach builds a predictive model for both classes and any new data instance is compared against these models. There are certain challenges in supervised outlier detection, such as outlier data instances are few as compared to normal data instances and even it is difficult to obtain the accurate class labels.

Semi-supervised Scenario

In some applications, only training data for the normal class (es) (or only the abnormal class (es)) are provided. In semi-supervised outlier detection mode, training dataset is available only for normal class. Hence it is widely used than supervised mode. The new target instance is compared against this normal class and the data instances which do not satisfies this class are considered as an outlier. This mode is not used commonly as it is difficult to cover each abnormal behavior to generate normal class.

Unsupervised Scenario

These techniques are widely used as they do not require training data set. In most applications there are no training data available. This technique assumes that normal data instances are more frequent than outliers. The data instances which are frequent or closely related are considered as normal instances and remaining are considered as outliers.

1.3 Steps to calculate outlier

An anomaly is a data point that is significantly different from the other data points in a sample data set. The term is used in statistical studies, and can point to abnormalities in the data set studied or errors in the measurements taken. By knowing how to calculate outliers is important for ensuring a proper understanding of the data, and will lead to more accurate conclusions drawn from the study. There is a straightforward process for calculating outliers in a given set of observations.

Step-1 Learn how to recognize an outlier.
Step-2 Arrange the data points from lowest to highest data value.
Step-3 Calculate the median of the data set.
Step-4 Calculate the lower quartile Q1
Step-5Calculate the upper quartile Q3
Step-6 Find the "inner fences" for the data set.Q3+ ((Q3-Q1)*1.5)
Step-7 Find the "outer fences" for the data set.Q3+ ((Q3-Q1)*3)

Any data points that lie outside the calculated range are considered mild anomaly and any data points that lie outside the outer fences are considered extreme anomaly.

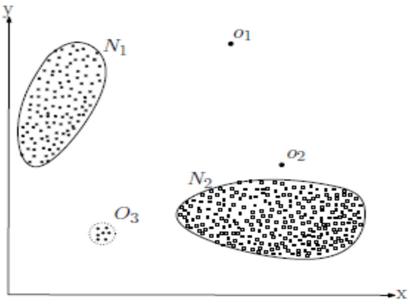

Fig. 1 Observation of outliers and normal data

1.4 Challenges

At abstract level, outlier can be defined as data that does not confirm the expected normal behavior. Simply outlier detection can be defined as a region which define or present region with normal behavior of data distribution & that does not represent are anomaly. But some factor affecting this simple definition is [8]:
-Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa.
-When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behavior more difficult.
- In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.
-The exact notion of an anomaly is different for different application domains. For example, in the medical domain a small deviation from normal (e.g. Fluctuations in body temperature) might be an anomaly, while similar deviation in the stock market domain (e.g. Fluctuations in the value of a stock) might be considered as normal. Thus applying a technique developed in one domain to another is not straightforward.
-Availability of labeled data for training/validation of models used by anomaly detection techniques is usually a major issue.
-Often the data contains noise which tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

Because of above problem outlier detection become more difficult in real life environment. Most of the application considers factors such as nature of data, availability of labeled data, and type of anomalies to be detected.

1.5 Applications

A more exhaustive list of applications that exploit outlier detection is provided below [4]
- 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.
-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 neo taxis 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- 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.

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 [4].

2. Literature survey

2.1. Distribution based or statistic based

The statistical approach assumes data follows some predefined distribution and aims to find out outlier which deviates from such distribution. The underlying principle of any statistical outlier detection technique is: “An outlier is an observation which is suspected of being partially or wholly irrelevant as it is not generated by the stochastic model assumed” [16]. Generally the data distribution is not known previously, especially for high dimensional data. The statistical outlier detection approach depends on the nature of statistical model that is required to be fitted on the data. The main problem with these techniques is that in a number of situations, the user might not have much knowledge about the underlying data distribution [5].

Statistical method first compute the parameters assuming all data points have been generated by statistical distribution (e.g., mean and standard deviation) and consider outliers are points that have a low probability to be generated by the overall distribution (e.g., deviate more than 3 times the standard deviation from the mean) .

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 [4].Example of this approach are Gaussian distribution and Multivariate.

2.2. Clustering based

Clustering based approach [6] always apply a clustering based method on sample of data to characterize the local behaviors’ of the data. The sub-clusters contain significantly less data points than remaining clusters, are termed as outliers. Most of the earlier clustering-based anomaly detection methods found outliers as the byproduct of a clustering. Hence any data point which does not comply with any cluster is called an outlier. As the main aim is to find clusters, these approaches are not optimized to find outliers. The advantage of the cluster based technique is that they do not have to be supervised. Moreover, clustering based techniques are capable of being used in an incremental mode i.e. after learning the clusters, new points can be inserted in to the system and tested for the outliers.

Clustering based approaches are computationally expensive as they compromise huge computation of pair wise distances. Clustering algorithms are optimized to find clusters rather than outliers. Accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters and set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers. The performance of outlier detection is limited because, the clustering based approaches are unsupervised and do not require labeled training data. Examples of this method are k means clustering algorithm and fuzzy c-means (FCM).

2.3. Distance based

This method Judge a point based on the distance(s) to its neighbors. Basic Assumption of this method are Normal data objects have a dense neighborhood and Outliers are far apart from their neighbors, i.e., have a less dense neighborhood .The notion of distance-based (DB) outlier is been defined [8]: *“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. 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 kth nearest neighbor. Example is Mahalanobis distance.

2.4. Density based

Another approach, density based approach [7] i.e. local outlier detection (LOF), determines degree of outlierness of each data instance based on its local density. 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 [10].This approach identifies the data structure via density estimation. Also, it compares the density around a point with the density around its local neighbors. The relative density of a point compared to its neighbors is computed as an outlier score. Approaches also differ in how to estimate density.

Basic assumptions of this approach are the density around a normal data object is similar to the density around its neighbors and the density around an outlier is considerably different to the density around its neighbors. The advantage of these approaches is that they do not need to make any assumption for the generative distribution of the data. In this approach we have to calculate the distance between each data instance and all other data instances, it causes high computational complexity. Examples of this approach are LOF and top-n algorithm.

2.5. Neural network

Neural network based outlier detection approaches works in two steps. In first step neural network is trained to build the normal classes. In second step each target instance is tested against those classes by providing input to neural network. If the neural network accepts the data instance then it is normal and if the neural network rejects data instance it is termed as an outlier [16]. Distinct types of neural networks are derived from basic neural network. Replicator neural network has invented for one class outlier detection. RNN degrades with datasets containing radial outliers .RNN performs satisfactory for small and large datasets.

Neural Network methods often have difficulty with such smaller datasets [17]. 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 [17].

2.6. Fuzzy logic

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. 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. Since fuzzy logic is built atop the structures of qualitative description used in everyday language, fuzzy logic is easy to use [14]. 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 [15]

2.7. Proximity based

It examine the spatial proximity of each object in the data space, if the proximity of an object considerably deviates from the proximity of other objects it is considered an outlier. For Deviation based approach given a set of data points (local group or global set) outliers are points that do not fit to the general characteristics of that set, i.e., the variance of the set is minimized when removing the outliers. Outliers are the outermost points of the data set. For distance based approach it judge a point based on the distance(s) to its neighbors, several variants proposed. It assumes that normal data objects have a dense neighborhood and outliers are far apart from their neighbors, i.e., has a less dense neighborhood.

2.8. Wavelet based

Outlier detection by means of the wavelet transform is a recent study area [11]. 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 signal-cutting problem is solved by the fully scalable modulated window. 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 [13]. The kernel function is called the *mother wavelet*, and the modifications are translations and compressions of the mother wavelet. 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 [12]. Example of it is Find Out method [11]

2.9. High dimensional approach

It examines the spectrum of pair wise angles between a given point and all other points. Outliers are points that have a spectrum featuring high fluctuation. The main challenge in data mining is curse of dimensionality. Relative contrast between distances decreases with increasing dimensionality. Data is very sparse, almost all points are outliers. Because of this concept of neighborhood becomes meaningless. Solutions to above problem is to use more robust distance functions and find full-dimensional outliers and find outliers in projections (subspaces) of the original feature space .Rational of this approach are angles are more stable than distances in high dimensional spaces (e.g. the popularity of cosine-based similarity measures for text data),object o is an outlier if most other objects are located in similar directions and object o is no outlier if many other objects are located in varying directions. Basic assumptions made are outliers are at the border of the data distribution and normal points are in the center of the data distribution. Example of it is angle based outlier detection.

2.10. Model based

Apply a model to represent normal data points and outliers are points that do not fit to that model. Model based approach uses a predictive model to characterized the normal data and then detect outlier as deviation from model [9].Most of the model based approaches implemented consider that input training data are perfectly labeled for building the outlier detection classifier or model. However the collected data may contaminated by noise and causes data with imperfect labels. Because of this imperfect label the normal data may behaves like abnormal data or outlier even though it may not be an outlier. This kind of result is known as uncertain data information and might cause labeling imperfection or errors into the training data, which further limits the accuracy of given outlier detection method .Example of it are svdd based and svm based outlier detection.

2.11. Support vector machine

Support Vector Machines (SVMs) [19] is a popular machine learning technique, which has been successfully applied to many real-world classification problems from various domains. Due to its theoretical and practical advantages, such as solid mathematical background, high generalization capability and ability to find global and non-linear classification solutions, SVMs have been very popular among the machine learning and data mining researchers. Although SVMs often work effectively with balanced datasets, they could produce sub optimal results with imbalanced datasets.

2.12. Support vector data description

Support vector data description (SVDD) by [20] is a method Support vector data description (SVDD), is a useful method for outlier detection. Its model is obtained by solving the dual optimization problem to find the boundary around a data set .SVDD constructs a little sphere around the normal data and use this sphere to detect unknown sample as outlier or normal one. SVDD transforms the original data into a feature space via a kernel function to detect outlier in high dimensional data. But its performance is affected by noise involved in the input data. Therefore, we will use SVDD method for outlier detection, which gives decision boundary around the normal data, and uses the few negative examples to refine the boundary to build an outlier detection classifier. The basic idea of SVDD method is to enclose all normal data example inside the sphere and exclude abnormal.

2.13 ABOD

In this section, we find all outlier models proposed so far inherently are not suitable for the requirements met in data mining high-dimensional data since they depends implicitly or explicitly on distances. This method not only use the distance between points in a vector space but primarily the directions of distance vectors. Comparing the angles between pairs of distance vectors to other points helps to discern between points similar to other points and outliers. The angle-based outlier factor (ABOF) is used to describe the divergence in directions of objects relatively to one another[18].ABOD calculates the variation of the angles between each target instance and the remaining data points, since it is observed that an outlier will produce a smaller angle variance than the normal ones do. Most outlier detection models require the user to specify parameters that are crucial to the outcome of the approach. For unsupervised approaches, such requirements are always a drawback. Thus, a big advantage of ABOD is being completely free of parameters [18].The concern of ABOD is the computation complexity due a huge amount of instance pairs to be considered.

2.14 Decremental PCA

In this method it calculates the principal direction the data. It is observed that removing (or adding) an abnormal data instance will affect the principal direction of the resulting data than removing (or adding) a normal one. Using this “leaves one out” (LOO) strategy, the principal direction of the data set without the target instance present and that of the original data set is calculated. Thus, the outlierness (or anomaly) of the data instance can be determined by the variation of the resulting principal directions. More incisively, the difference between these two eigenvectors will indicate the anomaly of the target instance. By ranking the difference outlier scores of all data points, one can easily identify the outlier data by a predefined threshold or a predetermined portion of the data. This framework can be considered as a decremental PCA (dPCA)-based approach for anomaly detection. The drawback of this method is it work for moderate size of data. In real-world anomaly detection a problem dealing with a large amount of data, removing/adding one target instance only creates negligible difference in the resulting eigenvectors, and the dPCA technique for anomaly detection fails to find outlier.

2.15 Oversampling PCA

To address the problem occurred in dPCA, osPCA is introduced. In this method the target instance is duplicated and then applies osPCA on such oversampled data set. By doing these effects of the outlier instance is amplified due to its duplicates present in the principal component analysis formulation and make outlier detection simple. However osPCA based outlier detection procedure with an oversampling strategy will markedly increase the computational load. For each target instance, it always needs to create a dense covariance matrix and solves the associated PCA found problem. This will prohibit the use of our proposed framework for real-world large-scale applications. Another power method is capable to produce approximated PCA solutions, but it requires the storage of the covariance matrix and it cannot be easily extended to applications with streaming data or online settings.

Proposed work

Most of the application considers data is perfectly classified in normal and abnormal classes. But labels of data get damaged due to noise or hardware error. This limits the performance of classifier .And negative data present in dataset in very small amount. This increase cost to detect outlier in such huge data.
To overcome above problem we introduce new approach, which consider object label and likelihood value i.e. degree of membership of object towards its own class for building a more model to detect outlier.

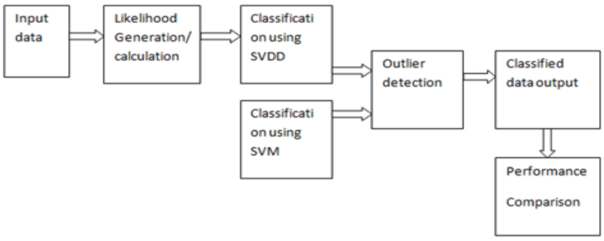

Fig.2. Proposed system for outlier detection

Outline of the approach

1. Cluster the Data through Kernel K-means clustering method
2. Calculate the Single-likelihood values i.e. (degree of membership towards its class (normal and abnormal))for each data object for every cluster.
3. Develop the Classifiers.
4. Test the dataset to measure performance.

This section provides a detailed description about our proposed approach to outlier detection. Outlier detection refers to the detection of a data set which is inconsistent with the remaining set of data. We will be going to use diabetes training dataset for outlier detection. This training dataset consists of normal examples and small amount of outlier (or abnormal) examples.

Our objective is to build a classifier which consider normal and abnormal training data and classify the unseen test data. In our project we are going to use support vector data description (SVDD) classifier. So the first step will be to generate pseudo training dataset by calculating likelihood values for each input data. We will use kernel k-means clustering algorithm to generate likelihood values for each input data. Afterwards we will apply SVDD on likelihood values and it will classify the test data into normal and abnormal class.

Conclusions

Outlier detection is an extremely important problem with direct application in a wide variety of domains. An important observation with outlier detection is that it is not a well-formulated problem. We have discussed the different ways in which the problem has been formulated in literature. Every unique problem formulation has a different approach, resulting in a large literature on outlier detection techniques. Several approaches have been proposed to target a particular application domain. The survey can hopefully allow mapping of such existing approaches to other application domains.

References

[1] Yu, D., Sheikholeslami. G. and Zang ,(2002) “A find out: finding outliers in very large datasets”, In Knowledge and Information Systems, pp.387 - 412.
[2] Breunig, M.M., Kriegel, H.P., and Ng, R.T.,(“LOF: Identifying density based local outliers.”, ACM Conf)erence Proceedings, pp. 93-104.
[3] Aggarwal, C. C., Yu, S. P., (2005) “An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, vol. 14,pp. 211–221.
[4] Hodge, V.J. (2004), *A survey of outlier detection methodologies*, Kluver Academic Publishers, Netherlands, January 2004.
[5] Xue, ()A.: Study on Spatial Outlier Mining. Zhen Jiang, Jiang Su University [6] S. Y. Jiang and Q. B. An. Clustering-based outlier detectiionmethod.*ICFSKD*, pages 429–433
[7] M. Breunig, H.-P.Kriegel, R.T. Ng, and J. Sander,(“LOF: Identifying Density-Based Local Outliers,” Proc.) ACM SIGMOD Int’l Conf. Management of Data.
[8] Knorr, E.M..; Ng, R. (1988). Algorithms for Mining Distance-Based Outliers in Large Datasets., *Proceedings of VLDB,* pp.392-403.
[9] C. Li and W. H. Wong, (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and

outlier detection. *In Proceedings of the National Academy of Sciences USA*, 98:31–36.

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

[11] 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.

[12] Yu, D.; Sheikholeslami G. & Zhang, A. (2002) Find Out: Finding Outliers in Very Large Datasets.
Knowledge and Information's Systems, vol.4, pp. 387-412, Springer-Verlag, London.

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

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

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

[16] V. Chandola, A. Banerjee, and V. Kumar, (2009) "Anomaly detection: A survey," ACM CSUR, vol. 41, no. 3, Article 15.

[17] 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*.

[18] H.-P. Kriegel, M. Schubert, and A. Zimek, (2004)"Angle-Based Outlier Detection in High-Dimensional Data," Proc. 14th ACM SIGKDD

[19] C. Cortes and V. Vapnik, (1995) "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297,.

[20]D. M. J. Tax and R. P. W. Duin,(2004) "Support vector data description," Machine Learning, vol. 54, no. 1, pp. 45–66.