CVME Seminar

Outlier Detection

Dominik Schörkhuber

Vienna University of Technology, May 2015

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Introduction

Outlier Detection is the identification of data which is not conform to an expected pattern. Unlike in clustering, where items of common behavior are grouped together, and

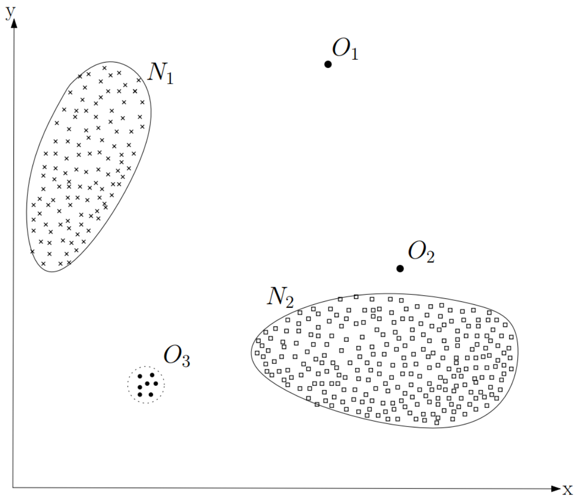

Abbildung 1: Example for Outlier Detection

outliers may be seen as unwanted noise, we are actively searching with outlier detection methods for anomalies rather than common patterns in data. Outlier detection is a heavily researched topic and has in its generality a wide range of applications. Some examples are fault detection in safety critical systems, supervision of banking systems, video surveillance or intrusion detection in cyber security. In all those cases we want to automatically distinguish usual behavior from abnormal behavior, which may indicate a system breakdown, fraud, etc. Outliers may also be of use in a more positive way. For example human resource management may use outlier detection techniques to find people capable of a very diverse set of skills. From those examples we can conclude that in many cases abnormal data may be of great interest instead of data following a common pattern. Until now the very generic term data was used, we now want to closer specify how input data for outlier detection algorithms may look like. A single data instance consists of multiple attributes, similar to descriptors in image analysis. Data attributes can be classified in different types of data like binary, categorial, discrete or continous data. Each data instance may contain one (univariate) or multiple (multivariate) attributes. In the multivariate case the attribute types may be coherent or different among attributes. Consider Fig. 1 as an abstract example with input data in the attribute space $\mathbb{R}^2$. Each data point consists of two attributes $p = (x, y)$ with $x, y \in \mathbb{R}$. The data is split into several regions. N_1 and N_2 refer to normal regions, those points describe normal behavior. O_1 and O_2 are single outliers, O_3 is an outlying region.

1.1 Types of Supervision

Outlier detection methods can be classified by their type of supervision. Besides the actual data instances an outlier detection algorithm may also use additional information for distinction of outliers from inliers. Such information can be class labels as they are used in pattern matching algorithms. The data instances of a training dataset is augumented with class information. This information enables us to generate a predictive model which may classify data instances of test data. Depending on how much a method utilizes this additional information we distinguish three classes of supervision.

Supervised methods make heavy use of labeling information. These techniques require class (labeling) information for each normal and outlier point in the training dataset. Typically a predicitive model is built which enables the algorithm to classify further data points into normals or outliers. The main drawback of supervised methods is it may be very expensive to acquire correctly labeled training datasets. Depending on the field of use a human expert may be required to do a correct labeling or atleast correct a given automatic labeling of training data.

Semi-supervised methods on the other hand only require labeling information for one class of points. Either only normal or abnormal behavior is captured by the training set. It is often very difficult to provide correct labels for both classes of points. For example in a network intrusion detection system it is usually impossible to model all possible attack vectors as outliers in a training dataset. Whereas information of normal use of the network may be automatically acquired. For this reason unsupervised methods with known sets of outliers are not very popular. Since normal behavior is easier to model techniques with known normal datasets are usually used.

Unsupervised methods as a third method does not rely on any additional information. Therefore these methods are widely applicable since no additional information must be provided. Still we need some way of categorising input data points. For example statistic methods can be used to adapt a parametric distribution to normal and/or abnormal data. Based on a statistical test we may distinguish inliers from outliers. Also many techniques rely on the fact that normal data instances are occuring more frequently than abnormal data instances. Therefore frequently occuring patterns are classified as normal, whereas rare patterns are assumed to be abnormal.

1.2 Results of Outlier Techniques

Another important aspect of an outlier detection technique is the manner in which an outlier is reported. In section 1 we refered to training data as labeled input data. Labeled either as normal data instance or as an outlier. For outputs of outlier detection techniques we distinguish two classes. The first class are labeling outlier techniques. Those techniques put binary labels on each point identifying them as outliers or inliers, just as we did before for training data. Often these techniques are also called hard classifiers [3]. On the other hand there are scoring outlier detection techniques, which are also called soft classification techniques. Scoring techniques determine an outlier score which indicates the outlierness of a point. Which is a continuous value ranging e.g. from being an inlier (0) to being an outlier(1). Scoring techniques enable us to create a ranking among data instances. Through this ranking we can find the most outlying point. Based on a threshold we can also compute binary labels for results computed by a scoring approach.

global vs local labeling vs scoring = hard / soft TypeI TypeII outliers

side products of clustering categories of outlier detection methods

1.3 Overview of Outlier Detection Techniques

proximity based angle based

2 Outlier Detection in High Dimensional Data

problems rising with high dimensionality

3 Application of Outlier Detection Techniques

4 Identifying Density-based Local Outliers

[1] [2] [3] [4] [5] [6]

Literatur

[1] Charu C Aggarwal and Philip S Yu. Outlier detection for high dimensional data. In *ACM Sigmod Record*, volume 30, pages 37–46. ACM, 2001.

[2] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *Principles of data mining and knowledge discovery*, pages 15–27. Springer, 2002.

[3] Irad Ben-Gal. Outlier detection. In *Data Mining and Knowledge Discovery Handbook*, pages 131–146. Springer, 2005.

[4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Outlier detection: A survey. *ACM Computing Surveys*, 2007.

[6] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.