Chapter 7

OUTLIER DETECTION

Irad Ben-Gal
Department of Industrial Engineering
Tel-Aviv University
Ramat-Aviv, Tel-Aviv 69978, Israel.
bengal@eng.tau.ac.il

Abstract Outlier detection is a primary step in many data-mining applications. We present several methods for outlier detection, while distinguishing between univariate vs. multivariate techniques and parametric vs. nonparametric procedures. In presence of outliers, special attention should be taken to assure the robustness of the used estimators. Outlier detection for Data Mining is often based on distance measures, clustering and spatial methods.

Keywords: Outliers, Distance measures, Statistical Process Control, Spatial data

1. Introduction: Motivation, Definitions and Applications

In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis (Williams *et al.*, 2002; Liu *et al.*, 2004).

An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins (1980) defines an outlier *as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.* Barnett and Lewis (1994) indicate that *an outlying observation, or outlier, is one that appears to deviate markedly from other*

members of the sample in which it occurs, similarly, Johnson (1992) defines an outlier *as an observation in a data set which appears to be inconsistent with the remainder of that set of data.* Other case-specific definitions are given below.

Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks (Hawkins, 1980; Barnett and Lewis, 1994; Ruts and Rousseeuw, 1996; Fawcett and Provost, 1997; Johnson *et al.*, 1998; Penny and Jolliffe, 2001; Acuna and Rodriguez, 2004; Lu *et al.*, 2003).

2. Taxonomy of Outlier Detection Methods

Outlier detection methods can be divided between *univariate methods*, proposed in earlier works in this field, and *multivariate methods* that usually form most of the current body of research. Another fundamental taxonomy of outlier detection methods is between *parametric* (*statistical*) methods and *non-parametric* methods that are model-free (e.g., see (Williams *et al.*, 2002)). Statistical parametric methods either assume a known underlying distribution of the observations (e.g., (Hawkins, 1980; Rousseeuw and Leory, 1987; Barnett and Lewis, 1994)) or, at least, they are based on statistical estimates of unknown distribution parameters (Hadi, 1992; Caussinus and Roiz, 1990). These methods flag as outliers those observations that deviate from the model assumptions. They are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution (Papadimitriou *et al.*, 2002).

Within the class of non-parametric outlier detection methods one can set apart the data-mining methods, also called *distance-based methods*. These methods are usually based on local distance measures and are capable of handling large databases (Knorr and Ng, 1997; Knorr and Ng, 1998; Fawcett and Provost, 1997; Williams and Huang, 1997; Mouchel and Schonlau, 1998; Knorr *et al.*, 2000; Knorr *et al.*, 2001; Jin *et al.*, 2001; Breunig *et al.*, 2000; Williams *et al.*, 2002; Hawkins *et al.*, 2002; Bay and Schwabacher, 2003). Another class of outlier detection methods is founded on *clustering techniques*, where a cluster of small sizes can be considered as clustered outliers (Kaufman and Rousseeuw, 1990; Ng and Han, 1994; Ramaswamy *et al.*, 2000; Barbara and Chen, 2000; Shekhar and Chawla, 2002; Shekhar and Lu, 2001; Shekhar and Lu, 2002; Acuna and Rodriguez, 2004). Hu and Sung (2003) , whom proposed a method to identify both high and low density pattern clustering, further partition this class to *hard classifiers* and *soft classifiers*. The former partition the data into two non-overlapping sets: outliers and non-outliers. The latter offers a ranking by assigning each datum an outlier clas-