Clustering Techniques for Outlier Detection

Frank Klawonn
University of Applied Sciences Braunschweig/Wolfenbuettel, Germany

Frank Rehm
German Aerospace Center, Germany

INTRODUCTION

For many applications in knowledge discovery in databases, finding outliers, which are rare events, is of importance. Outliers are observations that deviate significantly from the rest of the data, so they seem to have been generated by another process (Hawkins, 1980). Such outlier objects often contain information about an untypical behaviour of the system.

However, outliers bias the results of many data-mining methods such as the mean value, the standard deviation, or the positions of the prototypes of k-means clustering (Estivill-Castro & Yang, 2004; Keller, 2000). Therefore, before further analysis or processing of data is carried out with more sophisticated data-mining techniques, identifying outliers is a crucial step. Usually, data objects are considered as outliers when they occur in a region of extremely low data density.

Many clustering techniques that deal with noisy data and can identify outliers, such as possibilistic clustering (PCM) (Krishnapuram & Keller, 1993, 1996) and noise clustering (NC) (Dave, 1991; Dave & Krishnapuram, 1997), need good initializations or suffer from lack of adaptability to different cluster sizes. Distance-based approaches (Knorr & Ng, 1998; Knorr, Ng, & Tucakov, 2000) have a global view on the data set. These algorithms can hardly treat data sets that contain regions with different data density (Breuning, Kriegel, Ng, & Sander, 2000).

In this work, we present an approach that combines a fuzzy clustering algorithm (Höppner, Klawonn, Kruse, & Runkler, 1999) or any other prototype-based clustering algorithm with statistical distribution-based outlier detection.

BACKGROUND

Prototype-based clustering algorithms approximate a feature space by means of an appropriate number of prototype vectors, where each vector is located in the centre of the group of data (the cluster) that belongs to the respective prototype. Clustering usually aims at partitioning a data set into groups or clusters of data, where data assigned to the same cluster are similar and data from different clusters are dissimilar. With this partitioning concept in mind, an important aspect of typical applications of cluster analysis is the identification of the number of clusters in a data set. However, when we are interested in identifying outliers, the exact number of clusters is irrelevant (Georgieva & Klawonn).

The idea of whether one prototype covers two or more data clusters or whether two or more prototypes compete for the same data cluster is not important as long as the actual outliers are identified and assigned to a proper cluster. The number of prototypes used for clustering depends, of course, on the number of expected clusters but also on the distance measure respectively the shape of the expected clusters. Because this information is usually not available, the Euclidean distance measure is often recommended with rather copious prototypes.

One of the most referred statistical tests for outlier detection is the Grubbs’ test (Grubbs, 1969). This test is used to detect outliers in a univariate data set. Grubbs’ test detects one outlier at a time. This outlier is removed from the data set, and the test is iterated until no outliers are detected.

MAIN THRUST

The detection of outliers that we propose in this work is a modified version of the one proposed in Santos-Pereira and Pires (2002) and is composed of two different techniques. In the first step we partition the data set with the fuzzy c-means clustering algorithm so the feature space is approximated with an adequate number of prototypes. The prototypes will be placed in the centre of regions with a high density of feature vectors. Because outliers are far away from the typical data, they influence the placing of the prototypes.

After partitioning the data, only the feature vectors belonging to each single cluster are considered for the detection of outliers. For each attribute of the feature vectors of the considered cluster, the mean value and the standard deviation has to be calculated. For the vector
Related Content

On Interactive Data Mining
www.igi-global.com/chapter/interactive-data-mining/10956?camid=4v1a

Microarray Databases for Biotechnology
www.igi-global.com/chapter/microarray-databases-biotechnology/10694?camid=4v1a

Entity Resolution on Names
www.igi-global.com/chapter/entity-resolution-on-names/103243?camid=4v1a

User-Centered Interactive Data Mining
www.igi-global.com/chapter/user-centered-interactive-data-mining/7748?camid=4v1a