Cluster Analysis 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, rare events, is of importance. Outliers are observations, which deviate significantly from the rest of the data, so that it seems they are generated by another process (Hawkins, 1980). Such outlier objects often contain information about an untypical behavior of the system.

However, outliers bias the results of many data mining methods like the mean value, the standard deviation or the positions of the prototypes of *k-means* clustering (Estivill-Castro, 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.

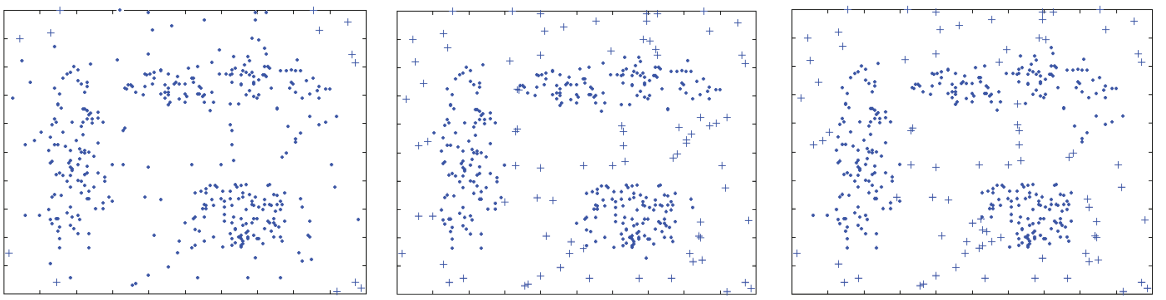
Many clustering techniques like possibilistic clustering (PCM) (Krishnapuram & Keller, 1993; Krishnapuram & Keller, 1996) or noise clustering (NC) (Dave, 1991; Dave & Krishnapuram, 1997) that deal with noisy data and can identify outliers, need good initializations or suffer from lack of adaptability to different cluster sizes (Rehm, Klawonn & Kruse, 2007). Distance-based approaches (Knorr, 1998; Knorr, Ng & Tucakov, 2000) have a global view on the data set. These algorithms can hardly treat data sets containing 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 prototype vector is located in the center 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, in typical applications of cluster analysis an important aspect 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, 2006). If one prototype covers two or more data clusters or if two or more prototypes

Figure 1. Outlier detection with different number of prototypes



(a) 1 prototype (b) 3 prototypes (c) 10 prototypes

compete for the same data cluster, this is not important as long as the actual outliers are identified and note 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. Since this information is usually not available, it is often recommended to use the Euclidean distance measure 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.

The detection of outliers as we propose in this work is a modified version of the one proposed in (Santos-Pereira & 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 that the feature space is approximated with an adequate number of prototypes. The prototypes will be placed in the center of regions with a high density of feature vectors. Since 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 with the largest distance[a] to the mean vector, which is assumed to be an outlier, the value of the z-transformation for each of its components is compared to a critical value. If one of these values is higher than the respective critical value, than this vector is declared as an outlier. One can use the Mahalanobis distance as in (Santos-Pereira & Pires, 2002), however since simple clustering techniques like the (fuzzy) c-means algorithm tend to spherical clusters, we apply a modified version of Grubbs' test, not assuming correlated attributes within a cluster.

The critical value is a parameter that must be set for each attribute depending on the specific definition of an outlier. One typical criterion can be the maximum number of outliers with respect to the amount of data (Klawonn, 2004). Eventually, large critical values lead to smaller numbers of outliers and small critical values lead to very compact clusters. Note that the critical value is set for each attribute separately. This leads to an axes-parallel view of the data, which in cases of axes-parallel clusters leads to a better outlier detection than the (hyper)-spherical view on the data.

If an outlier was found, the feature vector has to be removed from the data set. With the new data set, the mean value and the standard deviation have to be calculated again for each attribute. With the vector that has the largest distance to the new center vector, the outlier test will be repeated by checking the critical values. This procedure will be repeated until no outlier will be found anymore. The other clusters are treated in the same way.

Results

Figure 1 shows the results of the proposed algorithm on an illustrative example. The crosses in this figure are feature vectors, which are recognized as outliers. As expected, only few points are declared as outliers, when approximating the feature space with only one prototype. The prototype will be placed in the center

Table 1. Estimated flight duration before and after outlier treatment

Cluster	mean flight duration (s) (before outlier test)	RMSE	mean flight duration (s) (after outlier test)	RMSE
1	2021.18	266.17	2021.18	266.17
2	2497.13	407.90	2465.71	358.68
3	2136.85	268.93	2136.85	268.93
4	2303.41	409.35	2303.41	409.35
5	2186.22	292.04	2186.22	292.04
6	1872.23	180.45	1872.23	180.45
7	2033.31	395.33	2033.31	395.33
8	1879.28	187.12	1879.28	187.12
9	1839.65	90.95	1839.65	90.95
10	2566.15	517.01	2523.28	492.60