Outlier detection in astronomical data
Yanxia Zhang, Ali Luo, Yongheng ZHAO
National Astronomical Observatories, Chinese Academy of Sciences

ABSTRACT
Astronomical data sets have experienced an unprecedented and continuing growth in the volume, quality, and complexity over the past few years, driven by the advances in telescopes, detector, and computer technology. Like many other fields, astronomy has become a very data rich science. Information content measured in multiple Terasizes, and even larger, multi Petabytes data sets are on the horizon. To cope with this data flood, Virtual Observatory (VO) federates data archives and services representing a new information infrastructure for astronomy of the 21st century and provides the platform to science discovery. Data mining promises to both make the scientific utilization of these data sets more effective and more complete, and to open completely new avenues of astronomical research. Technological problems range from the issues of database design and federation, to data mining and advanced visualization, leading to a new toolkit for astronomical research. This is similar to challenges encountered in other data intensive fields today. Outlier detection is of great importance as one of knowledge discovery tasks. The identification of outliers can often lead to the discovery of truly unexpected knowledge in various fields. Especially in astronomy, the great interest of astronomers is to discover unusual, rare or unknown types of astronomical objects or phenomena. The outlier detection approaches in large datasets correctly meet the need of astronomers. In this paper we provide an overview of some techniques for automated identification of outliers in astronomical data. Outliers often provide useful information. Their identification is important not only for improving the analysis but also for indicating anomalies which may require further investigation. The technique may be used in the process of data preparation and also be used for preselecting special object candidates.

Keywords: Outlier-Data Mining-Data Mining Applications-Algorithms-Exceptions

DEFINITION
Outlier: it is defined as a data point which is very different from the rest of the data based on some measure. The points are neither apart of a cluster nor a part of the background noise; rather they are specifically points which behave very differently from the norm.

Possible sources of outliers:
• Data entry errors (recording and measurement errors)
• Incorrect distribution assumption, unknown data structure,
• Skewed data (e.g., the presence of non-valid data)
• Sampling errors
• Outlier is the point or elements being detected.

Effects:
• In astronomy, no finding outliers is equal to finding no new objects.

How to find outliers:
Methods for univariate outliers include z-Scores, box plot, histogram, and Barnett and Lewis (1994) provide a comprehensive treatment, listing about 100 discordancy tests for normal, exponential, and Poisson distributions. The choice of appropriate discordancy tests depend on:
(i) the distribution
(ii) whether or not the distribution parameters (e.g., mean and variance) are known
(iii) the number of expected outliers, and even
(iv) the type of expected outlier (e.g., lower or upper outliers in ordered sample).

Measures of Relative Location and Locating Outliers:
• z-Scores
• An outlier is an unusually small or unusually large data value in a data set.
• A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
• It might be an incorrectly recorded data value.
• It might be a data value that was incorrectly included in the data set.
• It might be a correctly recorded data value that belongs in the set.

Chebyshev’s Theorem:
At least (1 - 1/k^2) of the items in any data set will be within k standard deviations of the mean, where k is any value greater than 1.
At least 75% of the items must be within 2 = 2 standard deviations of the mean.
At least 99% of the items must be within k = 3 standard deviations of the mean.
At least 94% of the items must be within k = 4 standard deviations of the mean.

The Empirical Rule:
For data having a bell-shaped distribution:
• Approximately 68% of the data values will be within one standard deviation of the mean.
• Approximately 95% of the data values will be within two standard deviations of the mean.
• Approximately 99.7% of the data values will be within three standard deviations of the mean.

Detecting Outliers:
• An outlier is an unusually small or unusually large data value.
• A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
• It might be an incorrectly recorded data value.
• It might be a data value that was incorrectly included in the data set.
• It might be a correctly recorded data value that belongs in the data set.


useful to find univariate outliers:
All these methods of univariate outlier detection are based on unarguable order of data values. For N multivariate observations, there is no unambiguous total ordering. But different sub-orders have been suggested, of which the Mahalanobis distances (MDs) and the generalized distance metric (GDMs) are the most popularly used in the outlier study. We refer sub-ordering schemes to Barnett and Lewis (1994) for a relatively new technique to produce a set of scales P: 0 = 1 = 2 = N. It is produced by transforming each multivariate observation x into a scalar r. Then, it is sorted to produce the actual ordering of the multivariate data. The transformation is often done with a distance metric and, therefore, the extremes are those multivariate observations associated with the largest values of r. The sub-ordering used is based on the generalized distance metric:

r = \sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2 / \sigma_i^2}

where r is the sample mean vector and S is the sample covariance matrix.

For the ordered reduced univariate measures, we may adopt univariate outlier detection methods which are partially based in search in data. Measures of relative location and locating outliers include z-Scores, Chebyshev’s Theorem, and the Empirical Rule.

Another way is statistical. For a level of significance, the critical value is given as:

z = \alpha / 2

where z is the sample mean vector and S is the sample covariance matrix.

In short, the points are neither part of a cluster nor a part of the background noise; rather they are specifically points which behave very differently from the norm.

Applications:
1: electronic commerce
2: credit card fraud
3: performance statistics
4: credit card fraud
5: electronic commerce
6: credit card fraud
7: performance statistics
8: credit card fraud
9: performance statistics
10: performance statistics

STOP

Email: zyxz@bao.ac.cn; lal@bao.ac.cn; yzhao@lamost.org
National Astronomical Observatories, Chinese Academy of Sciences