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 telescope, detector, and computer technology. Like many other fields, astronomy has become a very data rich science. Information content measured in multiple Terabytes, and even larger, multi Petabyte 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 four 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 multivariate 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 preprocessing and also be used for preselecting special object candidates..

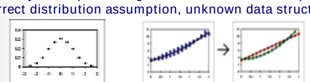
Keywords: Outlier-Data Mining-Data Mining Applications-Algorithms-Exceptions

DEFINATION

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 a part of a cluster nor a part of the background noise; rather they are specifically points which behave very differently from the norm. In short, *unusual data values.*

Possible sources of outliers

- Data entry errors(recording and measurement errors)
- Incorrect distribution assumption, unknown data structure,

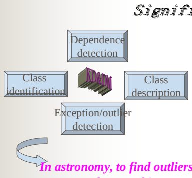


- Sometimes the cases are not a homogeneous set, but rather a heterogeneous set of two or more types of cases. One of these types will be far more frequent than the other, forcing the few to be identified as outliers.
- Rare events or novel phenomena

Effects

- means, variances, regression coefficients
- bias or distortion of estimates
- inflated sums of squares
- faulty conclusions

Significiance



Applications

1: electronic commerce
2: credit card fraud
3: performance statistics of professional athletes
4: and so on

In astronomy, to find outliers is equal to finding rare or unknown objects/ new phenomena

How to find outliers

Methods for univariate outliers include z-Scores, box plot, histogram, and so on. Barnett and Lewis (1994) provide a comprehensive treatment, listing about 100 discordancy tests for normal, exponential, Poisson, and binomial 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 outliers(e.g., lower or upper outliers in ordered sample). Most of the discordancy tests that we have encountered are univariate, and are specific to certain distributions having specific types and numbers of outliers. In numerous data mining situations where we do not know whether a particular attribute follows a normal distribution, a gamma distribution, and so on, we would have to perform extensive testing to find a distributions that fits the attribute. Furthermore, some of these tests may not be well-suited to large datasets.

methods to find univariate outliers

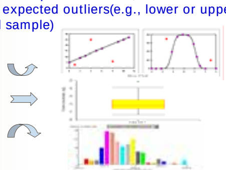
- Barnett and Lewis provide a comprehensive treatment, listing about 100 discordancy tests for normal, exponential, Poisson, and binomial distributions.

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 outliers(e.g., lower or upper outliers in ordered sample)

Scatter plot

Box plot

Histogram



•Measures of Relative Location and Locating Outliers

z -Scores

- An outlier is an unusually small or unusually large 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 data set!

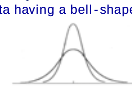
Chebyshev's Theorem

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 k = 2 standard deviations of the mean.
- At least 89% 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.
- Almost all of the items (99%) will be within three standard deviations of the mean.

Detecting Outliers

- An outlier is an unusually small or unusually large 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 data set!

methods to find multivariate 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-orderings have been suggested, of which the reduced sub-ordering is the most often used in the outlier study. Reduced sub-ordering is established in two phases. Firstly, a set of scalars $R=\{r_i\}(i=1,\ldots,N)$ is produced by transforming each multivariate observation x_i into a scalar r_i. Then, R 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 in R.

The sub-ordering used is based on the generalized distance metric:

$$r_i^2 = (x_i - x_0)\Gamma^{-1}(x_i - x_0)^T$$

where x_0 indicates the location of the data set and Γ^{-1} weights variables inversely to their scatter.Different choices of these parameters result in different distance metrics. For example, when Γ^{-1} is the identity matrix I, this equation defines the Euclidean distance of x_i to the location of the data set. When Mahalanobis distance is used in the multivariate outlier identification, the equation become as follows:

$$r_i^2 = (x_i - m)S^{-1}(x_i - m)^T$$

Where m is the sample mean vector and S is the sample covariance matrix.

For the ordered reduced univariate measures r_i, we may adopt univariate outlier detection method to evaluate whether outliers exist in data. Measures of relative location and locating outliers include z-Scores, Chebyshev's Theorem, and the Empirical Rule.

Another way is statistical technique. For α level of significance, the critical value is given

$$r_* = \frac{p(n-1)^2 F_{\alpha,p,n-p-1}}{n(n-p-1) + np F_{\alpha,p,n-p-1}}$$

where n is the sample size, p is the number of variables, and $F_{\alpha,p,n-p-1}$ is α level of significance of F-distribution with p and (n-p-1) degrees of freedom. If $r_i > r_*$, observation vector x_i is identified as an outlier at level α.

The studies on outlier detection can be broadly classified into six categories.

- The first is distribution-based, where a standard distribution (e.g. Normal, Poisson, etc.) is used to fit the data best and outliers deviate from the distribution
- The second category of outlier detection is depth-based which relies on the computation of different layers of k-d convex hulls.In depth-based methods, outliers are observations which distribution the outer layer of these hulls.
- A distance-based outlier in a dataset D is an object with pct% of the objects in D having a distance of more than d_{min} away from it.
- Another category is density-based, which applies to a certain degree to each object in a data set, depending on how isolated this object is, with respect to the surrounding clustering structure.
- The fifth category is clustering-based. Most clustering algorithms, especially those developed in the context of KDD (e.g. CLARANS, DBSCAN, BIRCH, STING, WaveCluster, DenClue, CLIQUE), are to some extent capable of handling exceptions
- Another kind of outlier detection is deviation-based. Genetic algorithms, which is an optimization technique based on various biological principles to detect outliers.
- Other kinds of outlier detection methods, such as fuzzy set theory, parallel algorithm, wavelet based multifractal formalism are in research.

VO-enabled data mining and knowledge discovery

Outlier detection is an important task for many KDD applications.In many proposals, outliers are only considered as a binary property. In this paper we show outlier detection is not a binary property, but a meaningful thing. Outlier detection in other fields is reviewed in detail, providing new thoughts and sights to detect outliers in astronomical data. Although this paper only skims the surface of dealing with outliers, it's presented with the hope that looking for unusual data values will become a regular part of our analysis, and that our research objectives and knowledge of our subject matter will help us decide what to do with them once we find them. A VO, federating various resources from different large, digital sky surveys, would enable us to thoroughly and systematically explore the observable space, and to understand the physical universe more completely and unbiasedly. Meanwhile, the VO may provide all kinds of data mining toolkits to mine the sky. We need develop efficient and effective outlier detection methods fit for characteristics of astronomical data in order to find the potential useful, rare or unknown types of objects and phenomena. These methods can be used to preselect source candidates and improve the efficiency of high-cost telescopes and enrich the data mining toolkits of VO.

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