

# Outlier detection in astronomical data

Yanxia Zhang<sup>\*a</sup>, Ali Luo<sup>\*a</sup>, and Yongheng Zhao<sup>\*a</sup>

<sup>\*a</sup>National Astronomical Observatories, Chinese Academy of Sciences, China.

## 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

## 1. INTRODUCTION

Huge quantities of data generated by modern astronomical instruments are collected and stored in databases, increasing the need for efficient and effective automated analysis methods to process the data and mining the information contained implicitly in the data. Knowledge discovery in databases (KDD) has been defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, et al. 1996). Corresponding to the kind of patterns to be discovered, several KDD tasks can be distinguished.

In general, KDD tasks can be classified into four categories: (a) dependency detection, (b) class identification, (c) class description, and (d) exception/outlier detection. The first three categories of tasks correspond to patterns that apply to many objects, or to a large percentage of objects, in the data. Most research in data mining, such as association rules, classification, and data clustering, belongs to these three categories. The fourth category, in contrast, focuses on a very small percentage of data objects, which are often ignored or discarded as noise. For example, some existing algorithms in machine learning and data mining have considered outliers, but only to the extent of tolerating them in whatever the algorithms are supposed to do.

There are many definitions for outliers which differ in words (Hawkins, 1980; Beckman & Cook, 1983; Barnett & Lewis, 1987; Tabachnick & Fidell, 1996). We use the one of Hawkins (1980), who defined an outlier is an observation that deviates so much from other observations so that it aroused suspicions that it is generated by a different mechanism. From a knowledge discovery standpoint, the rare events are often more

---

\*contact: zyx@lamost.bao.ac.cn; lal@lamost.bao.ac.cn; yzhao@lamost.bao.ac.cn; phone 8610-64841693; fax 8610-64878240; Datun Road 20A, Chao Yang District, National Astronomical Observatories, Chinese Academy of Sciences, China 100012

interesting than the common ones. For example, outlier detection has numerous applications, including credit card fraud detection, discovery of criminal activities in E-commerce, video surveillance, pharmaceutical research, weather prediction, the analysis of performance statistics of professional athletes and discovery of rare, unusual or unknown astronomical objects or phenomena. In addition, identification of outliers has recently begun to interest us for two reasons. Firstly, we consider balancing the imbalanced class distribution by reducing the largest classes before analysis. Outliers of the major classes seem to be worthwhile candidates for removal. Such outliers are treated as poor data which may be removed without further analysis. Secondly, the identification of outliers may give us some additional insight of data and often lead to the discovery of truly unexpected knowledge. Outliers are not outright dropped from data, instead they are presented to experts for further consideration.

In astronomy, systematic exploration of the observable parameter space, covered by large digital sky surveys spanning a range of wavelengths, will be one of the primary modes of research with a Virtual Observatory (VO) (Djorgovski, et al. 2001). The new, Vo-enabled astronomy is broadly grouped into two kinds. First, statistical astronomy will be performed, i.e. studies such as the mapping and quantification of the large scale structure in the universe, of the Galactic structure, construction and studies of complete samples of all kinds of objects. The second kind is to systematically explore the poorly known portions of the observable parameter space, and specifically search for rare, unusual, or even previously unknown types of astronomical objects and phenomena, e.g., as outliers in some parameter space of measured properties, both in the catalog and image domains. High-redshift quasars, type-2 quasars, brown dwarfs, and a small number of objects with puzzling spectra are examples of this. With the large numbers of detected sources, we may look for the rare events which would be unlikely to be found in smaller data sets. Rare objects may not be discriminated from the more common variety in some observable parameters, but be separated in other observable attributes.

In this paper, we first introduce the characteristics of astronomical data and the challenges by the complexity of the data. In detail, we review the sources and effects of outliers, and the techniques to detect outliers. In addition, we touch upon the data mining and knowledge discovery based on Virtual Observatory.

## 2. THE CHARACTERISTICS OF ASTRONOMICAL DATA

With the rapid development of observational technologies, astronomical data become more and more complex.

### 1. The huge volumes of astronomical data

Astronomical data are measured by Terabytes, even by Petabytes. Such huge amounts of data make some methods not to be applicable due to the difficulty of algorithms and huge efforts in computing. Therefore one task of knowledge discovery is to create new algorithm schemes and develop new efficient algorithms to overcome the technological difficulty caused by huge amounts of data.

### 2. The nonlinear relationship between the attributes of astronomical data

The nonlinear relationship between the attributes of astronomical data is the important characteristics of the complexness in the whole astronomical fields, of which there are complex mechanisms of inner interaction.

### 3. High dimensions of astronomical data

The next generation astronomy digital archives will cover most of the sky at fine resolution in many wavelengths, from X-rays through ultraviolet, optical, and infrared. In different wavelengths, astronomical data follow different laws and show different properties. A complete observable parameter space axes include quantities such as the object coordinates, velocities or redshifts, sometimes proper motions, fluxes at a range of wavelength (i.e., spectra; imaging in a set of bandpasses can be considered a form of a very low resolution spectroscopy), surface brightness and image morphological parameters for resolved sources, variability (or, more broadly, power spectra) over a range of time scales, etc. A typical VO data set may have the following properties:  $\sim 10^9$  data vectors in  $\sim 10^2$  dimensions. This is another complex characteristic of astronomical data. How to extract information from such high dimensional data is one challenge of knowledge discovery.

### 4. Heterogenous property of astronomical data

The data derived from different sources (e.g. from different surveys or from different projects) have different structures. Moreover the formats of data are various, for example, structural, half-structural and non-structural. The kinds of data include photometric data, spectral data, morphological data, image data, time data and so on.

#### 5. Missing values in astronomical data

The missing values arise from some unexpected outside force (e.g. low sensitivity of instruments, bad weather). Some objects are undetected in one or more bandpasses. So they lose the information of the bandpasses. How to restore the missing values and evaluate the intrinsic distribution parameters is one of difficulties to deal with the complexity of data.

All these challenges described above propose high requirement for data mining and knowledge discovery. At the same time, they are the strong push to accelerate the development of data mining and knowledge discovery.

### 3. THE SOURCES AND EFFECTS OF OUTLIERS

According to the number of data dimension, outliers are divided into univariate outliers and multivariate outliers. In some statistics literatures (White, 1992; Knorr, 2002), multivariate outliers are subdivided into two forms: gross outliers and structural outliers. Gross outliers are those observations that outlying for one or more individual attributes. In other words, a gross outlier is an outlier in one dimension for at least one variable. Structural outliers do not own this property. However, they are outlying relative to the covariance structure of the non-outlying data. Structural outliers may or may not be detected visually in 2-D scatter plots or 3-D spin plots. In some cases, they may be detected only when all  $k$  dimensions are simultaneously considered.

According to the distribution of the underlying data points, an outlier can be one or more of following entities (Knorr, 2002):

- a. an extreme or relatively extreme value
- b. a contaminant, that is, an observation from some other (possible unknown) distribution
- c. a legitimate, but surprising/unexpected data value
- d. a data value that was measured or recorded incorrectly

Human error often produces unintentional outliers. Data entry may be incorrect and missing value codes are sometimes used as real data. Outliers are frequently generated as the result of the natural variation of population or process one cannot control. These outliers are from the intended population, but their values are unusual in comparison with the normal values. It is also possible to have an outlier that is not a member of population due to a sampling error (Barnett & Lewis, 1987; Tabachnick & Fidell, 1996). In one word, various origins of outliers are summarized as follows :

- a. Data entry errors, due to recording and measurement errors, can produce outliers.
- b. Incorrect distribution assumption is given for unknown data structure.
- c. Sometimes the cases are not a homogeneous set to which a single model will apply, 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.
- d. The fourth cause of outliers is produced by error distributions with thick tails, in which extreme observations occur with greater frequency than expected for a normal distribution. Least squares solutions are fairly robust to violations of the assumption that the errors are normally distributed, except when the violation is that the distribution has thick tails. Ironically, sampling distributions that look quite different from a normal distribution cause little trouble, while these thick tail distributions raise some with F interpretations.
- e. Rare events or novel phenomena arise.

Developing techniques to detect outliers and understanding how they impact data analysis are extremely important parts of a thorough analysis, especially when statistical techniques are applied to the data. Due to outliers occurring in data, they cause different effects. For example, in the presence of outliers, any statistical test on sample means and variances can be distorted. Estimated regression coefficients that minimize the sum of squares for error are very sensitive to outliers. There are several other problematic effects of outliers, including

- a. bias or distortion of estimates
- b. inflated sums of squares (which make it unlikely you'll be able to partition sources of variation in the data into meaningful components.)
- c. distortion of p-values (statistical significance, or lack thereof, can be due to the presence of a few or even one unusual data value)
- d. faulty conclusions (it's quite possible to draw false conclusions if one hasn't looked for indications that there was anything unusual in the data)

Despite the difficulties, exploring why outliers exist can provide many clues to the development of better models. In fact, many great discoveries in human history can be traced to a researcher exploring some outlying or unusual value. Outliers may indicate that an important range of the data has been ignored that is worth knowing about. In astronomy, astronomers pay attention to the outlying data so that they discover high redshift quasars, brown dwarfs, pulsars, and so on. The exploration of observable parameter spaces, created by combining of large sky surveys over a range of wavelengths, will be one of the chief scientific purposes of a VO. This includes an exciting possibility of discovering some previously unknown types of astronomical objects or phenomena.

#### 4. METHODOLOGY

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.

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 (Barnett & Lewis, 1987; Barnett, 1976), of which the reduced sub-ordering is the most often used in the outlier study (Barnett & Lewis, 1987). Reduced sub-ordering is established in two phases (Barnett & Lewis, 1987; Barnett, 1976). Firstly, a set of scalars  $R = r_i (i = 1, \dots, N)$  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 (Barnett, 1976) 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 (Laurikkala, et al. 2000)

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

where  $x_0$  indicates the location of the data set and  $\Gamma^{-1}$  weights variables inversely to their scatter. Different choices of these parameters result in different distance metrics. For example, when  $\Gamma$  is the identity matrix  $I$ , (1) defines the Euclidean distance of  $x_i$  to the location of the data set.

Mahalanobis distance is used in the multivariate outlier identification, and obtained from (1) by selecting  $\Gamma$  to be the population covariance matrix  $\Sigma$  (Laurikkala, et al. 2000). In general, the population mean  $\mu$  was used as the location parameter (Barnett & Lewis, 1987; Tabachnick & Fidell, 1996; Gnanadesikan & Kettenring, 1972). Often the population values are unknown and they are estimated with sample mean vector  $m$  and sample covariance matrix  $S$  (Jain & Dubes, 1988; Boberg, 1999).

The estimate of the covariance matrix is as follows (Laurikkala, et al. 2000):

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

Mahalanobis distance incorporates the dependencies between the attributes (Laurikkala, et al. 2000). This property is essential in multivariate outlier identification, where the goal is to detect unusual value combinations. Many distance metrics, including Euclidean distance, utilize only location information and are, therefore, unsuitable for this task. Another advantage of Mahalanobis distance is that the unit of variable has no influence on the distance, because each variable is standardized to mean of zero and variance of one.

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. The  $z$ -score is often called the standardized value. It denotes the number of standard deviations a data value  $r_i$  is from the mean.

$$z_i = \frac{r_i - \bar{r}}{\sigma} \quad (3)$$

where  $\bar{r}$  and  $\sigma$  are the mean and variance of  $r_i (i = 1, \dots, N)$ , respectively. If  $|z_i| \geq 3$ ,  $x_i$  is considered as an outlier, otherwise,  $x_i$  is not an outlier.

Chebyshev's Theorem is described as: 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.

As Empirical Rule describes, 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.

Another way is statistical technique. For  $\alpha$  level of significance, the critical value is given

$$r_* = \frac{p(n-1)^2 F_{\alpha, p, n-p-1}}{n(n-p-1) + np F_{\alpha, p, n-p-1}} \quad (4)$$

where  $n$  is sample size,  $p$  is the number of variables, and  $F_{\alpha, p, n-p-1}$  is  $\alpha$ -level value 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  $\alpha$ .

In addition, the graphical methods, such as box plots or histograms, are used for visualization. For the box plot method, the box is defined by 3 long horizontal line segments which mark the lower quartile (Q1), median, and upper quartile (Q2); this covers 50% of the data. The range between the upper quartile and the lower quartile is defined as the interquartile range or IQ. The following quantities (called fences) are needed for identifying extreme values in the tails of the distribution:

1. lower inner fence:  $Q1 - 1.5IQ$
2. upper inner fence:  $Q2 + 1.5IQ$
3. lower outer fence:  $Q1 - 3.0IQ$
4. upper outer fence:  $Q2 + 3.0IQ$

A point beyond an inner fence on either side is considered a mild outlier. A point beyond an outer fence is considered an extreme outlier. Box plots are most suitable for exploring both symmetric and skewed quantitative data, but they can also identify infrequent values from categorical data (Laurikkala, et al. 2000).

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 (Barnett & Lewis, 1994). Over one hundred tests of this category, called discordancy tests, have been developed for different scenarios. A key drawback of this category of test is that most of the distributions used are univariate. There are some tests that are multivariate (e.g. multivariate normal outliers). But for many KDD applications, the underlying distribution is unknown. Fitting the data with standard distributions is costly, and may not produce satisfactory results. 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 distribute in the outer layer of these hulls. In theory, depth-based approaches could work

for large values of  $k$ . However, in practice, while there exist efficient algorithms for  $k = 2$  or  $3$  (Preparata & Shamos, 1988; Ruts & Rousseeuw, 1996; Johnson et al. 1998), depth-based approaches become inefficient for large data sets for  $k \geq 4$ . Knorr and Ng (1998) proposed the notion of distance-based outliers. 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. Their notion generalized many notions from distance-based approaches and is further extended by Ramaswamy et al. (2000), so that outliers can be more efficiently discovered and ranked. 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. Breuning et al. (1999, 2000) put forward an outlier detection approach named "OPTICS-OF", which is based on the same theoretical foundation as density-based cluster analysis. Their notion of an outlier is local in the sense that the outlier-degree of an object is determined by taking into account the clustering structure in a bounded neighborhood of the object. The merit of the method is more appropriate for detecting different types of outliers than previous approaches. 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 (Breuning, et al. 2000, and references therein). However, since the main objective of a clustering algorithm is to find clusters, they are developed to optimize clustering, and not to optimize outlier detection. The exceptions (called "noise" in the context of clustering) are typically just tolerated or ignored when producing the clustering result. Though the outliers are not ignored, the notions of outliers are essentially binary, and there are no quantification as to how outlying an object is. Another kind of outlier detection is deviation-based. Crawford et al. (1995) detect outliers using genetic algorithms, which is an optimization technique based on various biological principles. This approach is capable of searching for global optima among a vast number of choices. By intelligent but somewhat random generation of subsets of data, potential sets of outliers are identified by minimizing the residual sum squares produced by the least squares method. Other kinds of outlier detection methods, such as fuzzy set theory (Last & Kandel, 2001), parallel algorithm (Hung & Cheung, 2002), wavelet based multifractal formalism (Struzik & Siebes, 2000), are in research.

## 5. VO-ENABLED DATA MINING

Astronomers are accustomed to apply physical criteria for distinguishing their interested objects from the uninteresting majority. For example, most quasars have been found by the properties of radio-loud point sources and UV excess objects, etc. ; ultraluminous IRAS galaxies own anomalous large FIR/visible flux ratios; variable stars and distant supernovas have particular types of light curves; and so on. The discovery of quasars and radio-galaxies, or ULIRGs, or LMXBs, or intra-cluster x-ray gas depend on the fact that these objects discriminate themselves with unusual broad-band energy distributions.

Some objects are unresolved in some parameter space, but distinguishable in other parameter spaces. With their flux ratios in different bands, e.g. colors, some objects are founded. IRAS sources are classified as probable stars or galaxies by FIR flux ratios. Likewise, the color selection method is a good way to discover high- $z$  quasars. DPOSS discovered high- $z$  and type-2 quasars by the method (Djorgovski et al. 1998). Normal stars form a temperature sequence, as a banana-shaped locus of points in the parameter space of colors; while high- $z$  and type-2 quasars are away from the locus. This color parameter space provide a good discrimination among these types of objects. Similar method is also used to discover brown dwarfs in SDSS and 2MASS. A variant of this technique (based also on the Lyman-limit drop) is now applied to find galaxies at  $z \geq 3$  (Djorgovski et al. 2001 and references therein). With image morphology, we can classify and select galaxies of a particular type.

Though physical criteria and color selection, we find some rare but known types of objects. Certainly, looking for new kinds of objects is more exciting. In future, the most intriguing new prospect for a VO is the possibility of discovery of previous unknown types of astronomical objects and phenomena. A VO may provide a thorough, large-scale, unbiased, multi-wavelength census of the universe, which make it possible to find more interesting and meaningful objects or phenomena. While some may simply turn out to be little more than curiosities, others may be representative of genuine new astrophysical phenomena (Djorgovski et al. 2001). The whole process of data mining (DM) and knowledge discovery in databases (KDD) in a VO is illustrated in Figure 1. Much of the power of a VO will be the platform to access, rapidly and efficiently, multiple data archives across a broad range of wavelengths and provide federated data sets, which would enable new insight

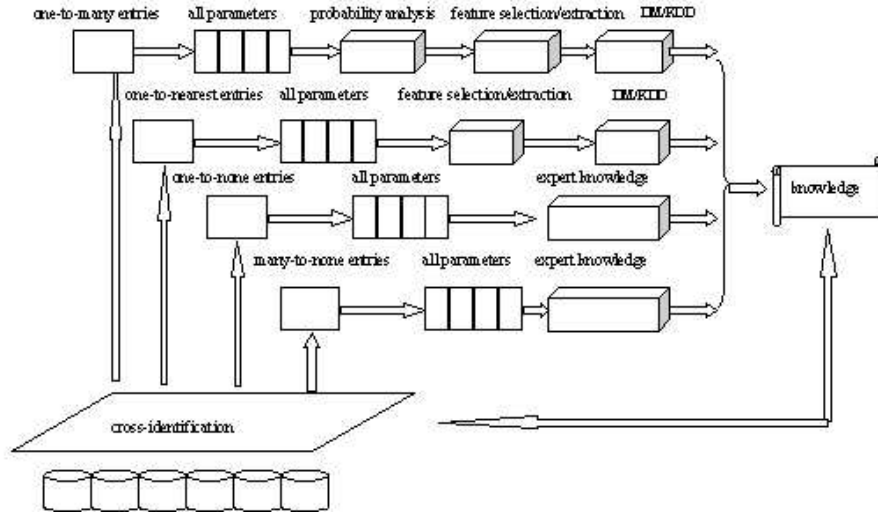


Figure 1. VO-enabled data mining and knowledge discovery

not obtained by any of data sets individually. Through multi-wavelength cross-identification, the VO federates various databases. The results are grouped into four forms: one-to-nearest entries, one-to-many entries, one-to-none entries, many-to-none entries. Such an example can refer to Zhang & Zhao (2003). For one-to-nearest entries, we deal with the attributes by feature selection or feature extraction and get optimized parameters used for DM/KDD. For one-to-many entries, probability analysis is needed to evaluate the probability of cross-identification and determine which is the true counterparts. For the last two situations, expert knowledge is necessary. Usually the two situations are especially important, for some unusual discoveries exactly arise from the situations. By means of DM/KDD and expert knowledge, new knowledge is obtained, then the knowledge is returned into the whole process and push the process to go forward. With more and more knowledge continuously accumulating, the theories will be enriched and pushed to go forward. As a result, we understand the universe more deeply. Although the process appears to be conceptually straightforward, it do require vast computational effort including the detection and federation of different datasets, automated statistical analysis and novel visualization of the final products. The results of the process will still require significant interpretation by users. The main advantages of a VO are to release scientists from the burden of data management and manipulation to spend their most precious resource and to improve their efficiency.

With the development of the quantity and quality of data , we are in urgent need of automated tools of data mining and knowledge discovery. This is generally falling in the area of unsupervised clustering, classification and exception/outlier detection. If the number of object classes is known and training data set of representative objects is available, the problem reduces to supervised classification. The searches of known types of objects with predictable signatures in the parameter space can be cast in this way. If the number of object classes is unknown, the data themselves can be grouped into some clusters in some objective and statistically sound manner by the unsupervised classification. The objects belonging to much less populated clusters are of special interest. In order to find special objects as many as possible, the outlier detection methods are in great demand. Many successful experiences in other fields are summarized in section 4. We need to learn their methods and transform the methods according to the characteristics of astronomy.

## 6. CONCLUSION

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.

## ACKNOWLEDGMENTS

This paper is funded by National Natural Science Foundation of China under grant No.10273011.

## REFERENCES

1. A. K. Jain, & R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, New Jersey, 1988.
2. B. G. Tabachnick & L. S. Fidell, Using Multivariate Statistics, HarperCollins, New York, 1996.
3. D. Hawkins. Identification of Outliers. Chapman and Hall, London, 1980.
4. E. M. Knorr, R. T. Ng, "Algorithms for Mining distance-Based Outliers in Large Datasets", Proc. 24th Int. Conf. on Very Large Data Bases, New York, NY, pp. 392-403, 1998.
5. E. M. Knorr, "Outliers and Data Mining: Finding Exceptions in Data", Doctor's thesis, Dept. of Computer science, University of British Columbia, 2002.
6. E. Hung, D. W. Cheung, "Parallel Algorithm for Mining Outliers in Large Database", in Distributed and Parallel Database, Kluwer Academic Publisher, 12, 2002.
7. F. Preparata, M. Shamos, "Computational Geometry: an Introduction", Springer, 1988.
8. I. Ruts, P. Rousseeuw, "Computing Depth Contours of Bivariate Point Clouds, Journal of Computational Statistics and Data Analysis, **23**, pp. 153-168, 1996.
9. J. Laurikkala, M. Juhola and E. Kentala, "Informal identification of outliers in medical data", the fifth workshop on Intelligent Data Analysis in Medicine and Pharmacology, Berlin, pp. 20-24, 2000.
10. J. Boberg, "Cluster Analysis: A Mathematical Approach with Applications to Protein Structures", Academic dissertation, Turku Centre for Computer Science, Turku, Finland, 1999.
11. K. D. Crawford, R. L. Wainwright, D. J. Vasicek, "Detecting Multiple Outliers in Regression Data Using Genetic Algorithms", Proc. of the 1995 ACM/AIGAPP Symposium on Applied Computing, Nashville, TN, pp. 351-356, ACM Press, 1995.
12. M. M. Breuning, et al. "OPTICS-OF: Identifying Local Outliers", Proc. of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99), Prague, pp. 262-270, 1999.
13. M. M. Breuning, et al. "LOF: Identifying Density-Based Local Outliers", Proc. ACM SIGMOD, Int. Conf. on Management of Data, Dallas, TX, pp. 93-104, 2000.
14. M. Last, & A. Kandel, "Automated Detection of Outliers in Real-World Data", Proc. of the Second International Conference on Intelligent Technologies, Bangkok, Thailand, pp. 292-301, 2001.
15. R. J. Beckman & R. D. Cook, "Outlier", Technometrics, **25**, pp. 119-149, 1983.
16. R. Gnanadesikan, & J. R. Kettenring, "Robust estimates, residuals, and outlier detection with multire sponse data", Biometrics, **28**, pp. 81-124, 1972.
17. R. A. White, "The detection and testing of multivariate outliers", Master's thesis, Dept. of statistics, University of British Columbia, 1992.



18. S. G. Djorgovski, A. A. Mahabal, R. J. Brunner, et al. "Searches for Rare and New types of Objects", *Virtual Observatories of the Future*, eds. R. J. Brunner, S. G. Djorgovski, and A. S. Szalazy, ASP Conference Series, **225**, 2001.
19. S. G. Djorgovski, R. R. Gal, S. C. Odewahn, et al. In *Wide Field Surveys in Cosmology*, eds. S. Colombi et al., Gif sur Yvette: Eds. Frontiers, 1998.
20. S. Ramaswamy, R. Rastogi, & K. Shim, Efficient algorithms for mining outliers from large data sets. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, Dallas, Texas, 2000.
21. T. Johnson, I. Kwok, R. Ng, "Fast Computation of 2-Dimensional Depth Contours", *Proc. 4th Int. Conf. on Knowledge Discovery and data Mining*, New York, NY, AAAI Press, pp. 224-228, 1998.
22. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework", *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, portland, pp.82-88, 1996.
23. V. Barnett & T. Lewis, *Outlier in statistical Data*, John Wiley & Sons, Norwich, 2nd edn., 1987.
24. V. Barnett & T. Lewis, *Outlier in statistical Data*, John Wiley & Sons, 1994.
25. V. Barnett, "The ordering of multivariate data (with discussion)", *Journal of the Royal Statistical Society A*, **139**, pp. 318-354, 1976.
26. Z. R. Struzik, A. P. J. M. Siebes, "Outlier detection and localisation with wavelet based multifractal formalism", Report INS-R0008, ISSN 1386-3681, 2000.
27. Y.-X., Zhang & Y.-H., Zhao, "Classification in Multidimensional Parameter Space: Methods and Examples", *PASP*, **115**, pp. 1006-1018, 2003.