RAA 2018 Vol. **18** No. **6**, 73(8pp) doi: 10.1088/1674–4527/18/6/73

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

Clustering analysis of line indices for LAMOST spectra with AstroStat

Shu-Xin Chen[1,2], Wei-Min Sun[1] and Qi Yan[1]

[1] College of Science, Harbin Engineering University, Harbin 150009, China; *sunweimin@hrbeu.edu.cn*
[2] College of Mechanical and Electrical Engineering, Qiqihar University, Qiqihar 161006, China

Received 2018 February 3; accepted 2018 April 13

Abstract The application of data mining in astronomical surveys, such as the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) survey, provides an effective approach to automatically analyze a large amount of complex survey data. Unsupervised clustering could help astronomers find the associations and outliers in a big data set. In this paper, we employ the k-means method to perform clustering for the line index of LAMOST spectra with the powerful software AstroStat. Implementing the line index approach for analyzing astronomical spectra is an effective way to extract spectral features for low resolution spectra, which can represent the main spectral characteristics of stars. A total of 144 340 line indices for A type stars is analyzed through calculating their intra and inter distances between pairs of stars. For intra distance, we use the definition of Mahalanobis distance to explore the degree of clustering for each class, while for outlier detection, we define a local outlier factor for each spectrum. AstroStat furnishes a set of visualization tools for illustrating the analysis results. Checking the spectra detected as outliers, we find that most of them are problematic data and only a few correspond to rare astronomical objects. We show two examples of these outliers, a spectrum with abnormal continuum and a spectrum with emission lines. Our work demonstrates that line index clustering is a good method for examining data quality and identifying rare objects.

Key words: methods: data analysis — techniques: spectroscopic — AstroStat — LAMOST

1 INTRODUCTION

Due to the limitations associated with observational equipment, traditional astronomy and astrophysics research efforts are basically based on small samples. With the improvement of observational capabilities of telescopes, more and more multi-object surveys have been conducted. Generally, large sky area surveys including photometry and spectroscopic survey projects have been carried out, and collections of survey data have become bigger and bigger. A representative telescope is the LAMOST facility, which took the lead in the world for its efficiency in obtaining celestial spectra (Cui et al. 2012). The Galactic surveys undertaken by LAMOST have produced a large amount of stellar spectra (Luo et al. 2012, 2015), and there are more than 7 million spectra observed from Nov 2011 to June 2016; the Fourth Data Release (DR4) has been made available (*http://dr4.lamost.org*). The LAMOST spectral analysis pipeline uses templates for stellar classification according to the similarity with observed templates (Wei et al. 2014).

Template based classification is a supervised approach, and all data can be grouped into the class for which the template is labeled. However, this approach would result in some rare objects dropping into a presupposed class which could not be detected. Data mining methods can automatically analyze a large amount of data, revealing hidden, previously unknown and potentially valuable information (Liu et al. 2015). Clustering, outlier analysis and feature learning are examples of methods that can be applied to mining data and discovering new knowledge. A very useful data mining tool, AstroStat (Kembhavi et al. 2015), has been developed since 2009, and is employed in this work to deal with LAMOST stellar spectra. We consider the performance of applying AstroStat to clustering LAMOST data.

The topic of clustering aims at partitioning and aggregating unlabeled data, and revealing hidden patterns of the data in an unsupervised way from the perspective of machine learning. Clustering is also a crucial task in scientific data analysis and engineering in various disciplines. By using the process of aggregating similar data together, the unaggregated data are obviously less similar according to some measure. An unsupervised clustering algorithm does not require the step of learning in advance, but the data still need to be preprocessed. Due to the diversity of structures and features in astronomical data, unsupervised clustering algorithms are suitable for recognizing the inherent distribution of the data and the hidden knowledge pattern without requiring information on classification. By analyzing outliers, we can identify a few data with characteristic anomalous spectra from the survey data.

Feature extraction is the most important step in data mining investigations. As is well known, absorption and emission lines are important features in a spectrum. The line index system is defined as a powerful feature extraction tool, and a series of values for line indices represent a spectrum with specific physical characteristics. The line index is measured from the equivalent width by integrating. A line index integrates the total flux of a spectral line or a magnitude of the multi-band at different wavelengths in a spectrum. A widely used line index system is the Lick line index, which has been applied in dealing with many survey data such as SDSS. In this paper, the Lick line index is used in LAMOST spectral data to extract atmospheric parameters, since the line index is almost unaffected by flux calibration errors and redshifts, and does not need any extinction correction because of the definition of line index. The line index is calculated from the average flux value over a relatively large wavelength range and incorporates normalization of spectral energy distribution, so it has a higher signal to noise ratio (S/N) than the original spectrum. When the spectral resolution changes, the line index will not change, unlike line fitting by using a pseudo-continuous spectrum. In LAMOST data releases (Wei et al. 2014), the line indices used in our research are provided in the form of arrays on LAMOST's official web site. Each value is named by a spectral line, which indicates the specific integral flux of spectral lines in a spectrum.

This paper is organized as follows. In Section 2, we introduce the AstroStat software and the line index data from LAMOST. In Section 3, we describe the k-means method, intra-cluster correlation analysis with Mahalanobis distance and employing a local outlier factor (LOF) to detect outlier measurements with distance. In Section 4, we search for outliers through the data mining process including using Mahalanobis distance, LOF factor, etc. In Section 4, we analyze the clustering result by checking the distance of outliers and their spectra. Finally, we conclude in Section 5.

2 TOOLS AND DATA

2.1 AstroStat

The interface of AstroStat, based on Java, was developed by the Virtual Observatory-India (VOI) project and allows astronomers to apply both simple and sophisticated statistical routines on large datasets. It is an easy-to-use software tool for statistical analysis and visual description applicable to big data. Users can freely download the software package from *http://vo.iucaa.ernet.in/voi/VOStat.html*. AstroStat loads the VOPlot service and the currently active data file, and uses the file for many kinds of data visualization. Some points of interest can be noted directly in the plot, which provides the possibility to select these points for manual analysis. We utilize VOPlot in AstroStat to realize the data analysis with visualization tools when analyzing LAMOST data. AstroStat shows the toolbar and k-means partitioning primary interface at the top which allows the user to select columns from multiple files in Figure 1.

2.2 R Language and LAMOST FITS

Data mining algorithms in AstroStat are based on the R language. The R language came into existence as a free counterpart of the S statistical language developed by Bell Labs. Ross Ihaka and Robert Gentleman (Ihaka & Gentleman 1996) developed R with many users involved, which resulted in a very large number of contributions from the users. It has all the common tools needed for advanced statistics: classification, clustering, etc.

AstroStat is based on the R language, and since R is a versatile open-source system for statistics, it integrates multiple data analysis and visualization methods. We decided to use R as the preprocessing step for LAMOST data. It not only has powerful data analysis capabilities, but also can effectively simplify the data analysis process.

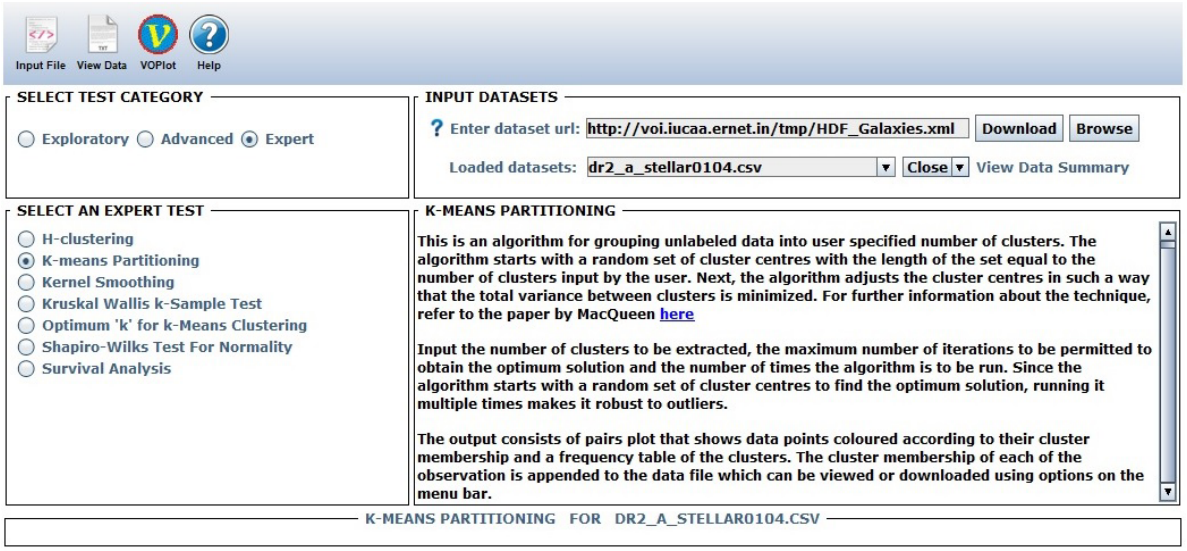

Fig. 1 A screenshot of AstroStat showing the toolbar at the top and the k-means partitioning primary interface which allows the user to select columns from multiple files.

The first step in preprocessing is reading the LAMOST FITS files by using the RFITSio package in R. Although AstroStat integrates the FITS reading package, we prefer to use the RFITSIO directly to avoid the problem of non-standard formatting in LAMOST spectra. The FITS data from LAMOST are written by the Cfitsio package, and the FITS file name is "spec-MMMMM-YYYY_spXX-FFF.fits" (Luo et al. 2012). Downloading the data and loading the required FITSio package with the function (FITSio), we can read the spectral data file by readFITS (" path * .fits "). Using the readFITS return value and the parameter, we can extract the eigenvector matrix and store it in format files "*.csv" and "*.txt", and then read the corresponding parameter information in the form of data columns and limit the maximum data range for values. We then use the function "plot()" to select type = "s" and select, for example, the A0III star spectral data file spec-55976-GAC_099N04_V5_sp12-128.fits in the LAMOST survey database, as shown in the plot of the flux spectrum in Figure 2.

2.3 Data

The LAMOST project processes spectra to have the same starting and ending wavelength scale at 3800Å to 9000Å respectively. In the data release, the Lick line indices are provided for A type stars. We download 144 340 A type stars with line indices released. An important goal in astronomy is to discover anomalous, sparse and even unknown types, and as an important contributing effort, outlier data mining can effectively identify a feature anomaly or the relatively tiny differences from the survey data from some day. Spectral lines are important main features in spectra, especially for A type stars because their continua are relatively smooth. The Lick line indices are extracted features of absorption lines which represent the physical character of a star.

3 CLUSTERING

As an important unsupervised classification method, cluster analysis is more suitable for investigating characteristics of data, and has been widely used in astronomical research. Qualitatively analyzing the clustering result is a key step for correlation analysis and outlier detection. By using the Mahalanobis distance as a metric for measurement, we do not need to consider the scale difference associated with line indices and it is easy to find spectra with small differences in features. In addition, for obvious outliers in a high dimensional space, the LOF factor with any distance measure could be quantitatively used for rare object detection.

3.1 K-means

K-means is a simple unsupervised learning algorithm that can solve classification problems as described by MacQueen et al. (1967). The given data are classified by finding the clusters and their related centers. Then each set of line indexes is assigned to the nearest center that

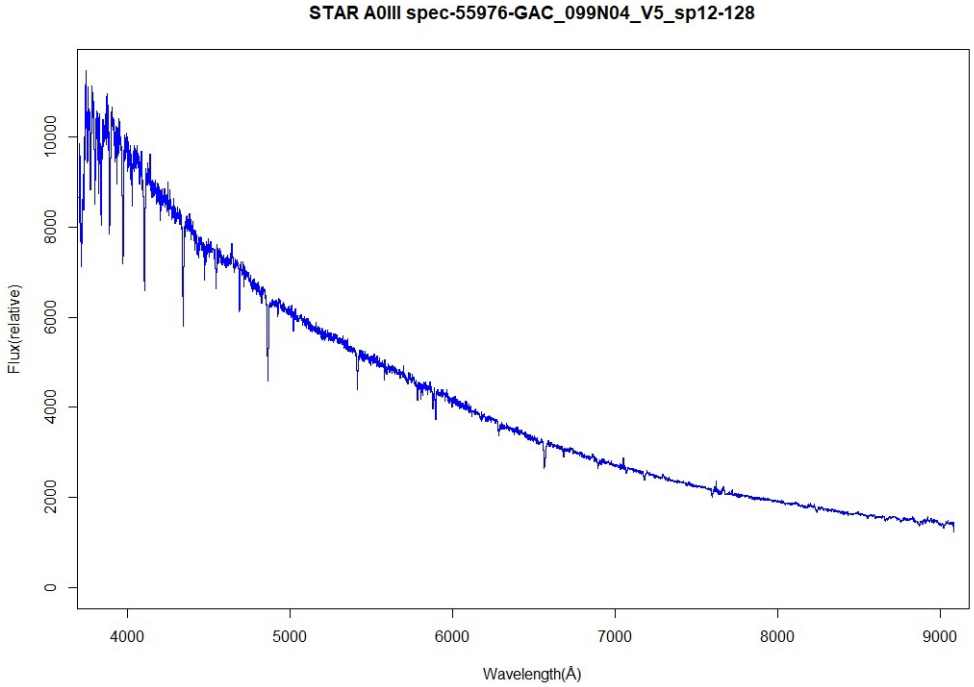

Fig. 2 A spectrum of an A0III-type star from the LAMOST survey data read and drawn by R.

is closest in a least squares sense. After all spectra are loaded, each point is replaced by the respective cluster center. Firstly, the algorithm initializes the cluster centers and normalizes the data. Theoretically, the initial cluster centers are found in the remaining data sets by using the maximum distance between every two pairs, excluding isolated points. Practically, the number of isolated points is often unpredictable, and the distances are calculated without excluding isolated points. Then, the two points with the largest distance are selected as the cluster centers for two classes. When the cluster centers are selected, multiple iterations are carried out, excluding the isolated points by checking if they exceed the threshold of a certain class. Finally, k-means assigns all the spectra to one of the clusters. The threshold range should be set to keep most data inside the two classes. The cluster centers of the k-means algorithm guarantee that objects in the same cluster are similar, while the objects in different clusters are not similar according to the specified metric.

The data in Figure 3 are from the LAMOST second data release (DR2) (*http://dr2.lamost.org*). Different line indices are clustered by the k-means algorithm using AstroStat, and Figure 3 illustrates that most A type stars are distributed in a small local area in the plane of kp12 vs. $H_\delta 12$. The clustering processing software platform in AstroStat is based on efficient, open source R language programming, and Figure 3 shows the result of clustering in the line index plane kp12 and $H_\delta 12$.

3.2 Intra-cluster Correlation Analysis with Mahalanobis Distance

In astronomical spectroscopy, similarity measures can be used to assess the closeness between the eigenvalues of arbitrary spectral lines. The k-means clustering algorithm in Figure 3 uses Euclidean distance to calculate the distance between two points. Mahalanobis distance uses the sample covariance method to measure the similarity of two unknown samples more effectively. Indian mathematician P.C. Mahalanobis first proposed Mahalanobis distance using sample covariance to calculate the distance between two points. For the set of m-dimensional points in Euclidean space defined by Equation (1), the Mahalanobis distance is $d_m(\boldsymbol{x}, \boldsymbol{y})$, where T is the transpose and $\boldsymbol{S}$ is the covariance matrix.

$$d_m(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}} \boldsymbol{S}^{-1} (\boldsymbol{x} - \boldsymbol{y})}. \quad (1)$$

Our experiment uses Mahalanobis (data, center = Avg, cov = $\boldsymbol{S}$) provided by R language, where Avg is the mean of the center and $\boldsymbol{S}$ is the sample covariance matrix. The function represents the Mahalanobis dis-

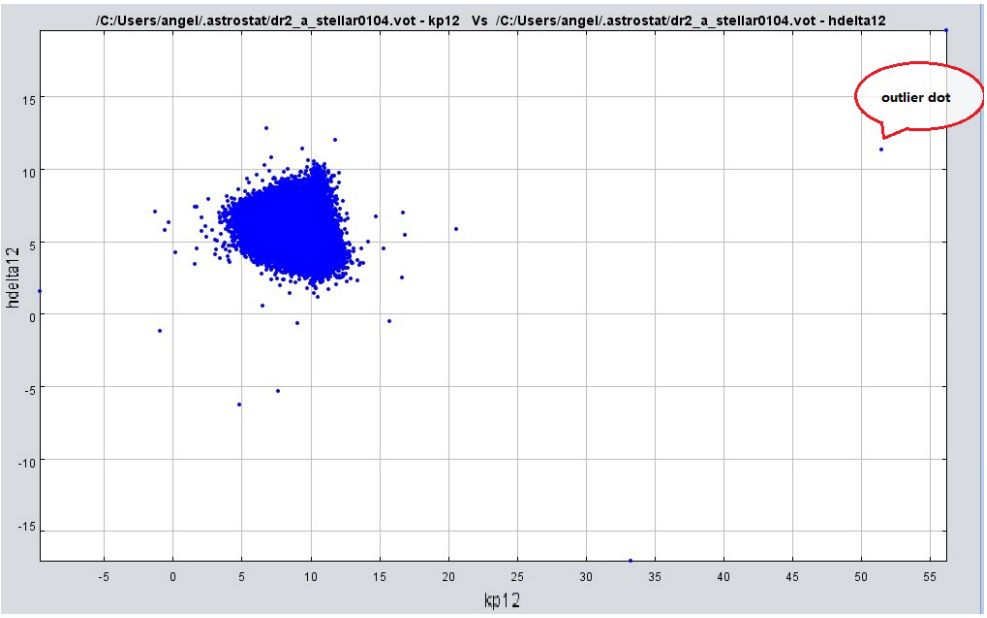

Fig. 3 These pictures show feature clustering in the dimension of kp12 and $H_\delta 12$ by AstroStat.

tance between each piece of data and the global library. The transformation of Mahalanobis distance is similar to that of the principal component analysis (PCA) solution, i.e., the PCA method rotates the principal component of the data to the x-axis in two-dimensional space, and scales it again to achieve the same measure of similarity. However, the Mahalanobis distance has no rotation transformation, and the similarity measure is scaled only in the x and y directions of the lower triangular inverse matrix. The Mahalanobis distance can take into account the relationships among various properties independently of the measurement scale. The definition of covariance matrix satisfies the four basic axioms of the spectral template distance: non-negativity, reflexivity, symmetry and trigonometric inequality. If the covariance matrix is a unit matrix, it is reduced to the Euclidean distance. We calculated Mahalanobis distance for each A-subtype of the dataset used in Figure 3, and the distances of median, maximum and minimum of each subtype are shown in Table 1.

In our process of experimental data processing, the Mahalanobis distance does not need to be normalized, but the Euclidean distance must be normalized first and then used to calculate the distance between points, otherwise the distance value is meaningless. Although the Euclidean distance is more commonly used, the extracted values of each spectrum represent different characteristics. Mahalanobis distance can better reflect the importance of different eigenvectors.

3.3 Outlier Measurement with Distance

LOF is used for outlier detection (Breunig et al. 2000), which can be applied to describe the singularity of normal A type spectra. For any positive integer k, the k-distance of object d, denoted as k-distance(d), is defined as the distance $d(d,o)$ between d and an object $o \in D$ such that Equation (2) is satisfied. LOF follows the definition of local reachable density; if a data point is far away from the distance between other points, it is clear that its local reachable density is small. A statistical anomaly detection algorithm usually needs to assume that the data follow a specific probability distribution. However, a clustering method usually only provides a judgment about whether or not the point is abnormal and cannot quantify the degree of abnormality for each point. In contrast, density-based LOF algorithms are simpler and more intuitive. The analysis does not require very much data that are part of a distribution to quantify outlierness. However, the LOF algorithm measures the degree to which a data point is anomalous not by looking at its absolute local density, but rather at the relative density of nearby data points. The benefit is that data can be distributed unevenly and with different densities. A local anomaly is defined by the relative local density. The lo-

Table 1 Analysis with Mahalanobis Distance for A Type Stars in Our Sample

Subclass	Count	Mean	Min	Max
A0	608	22.08346	6.217810	627.5468
A0III	951	35.60493	1.879113	380.4221
A1IV	18312	14.56612	2.270596	1452.920
A1V	18060	15.48386	4.545271	964.2106
A2IV	10970	14.91356	2.63847	966.1267
A2V	14195	17.20922	0.215421	3149.108
A3IV	2969	13.1970	0.839388	329.3868
A3V	596	21.61400	1.061798	1073.728
A5	16	960.2729	9.715015	10666.24
A5V	24421	11.00233	0.696124	5200.643
A6IV	14392	9.649854	0.671972	706.7618
A6V	1172	16.97445	1.513118	194.7997
A7III	1709	11.68790	0.226586	2560.394
A7IV	4873	9.291063	0.078180	894.8001
A7V	25159	7.853723	0.098206	2087.251
A8III	1882	6.725759	0.39664	116.1464
A9	33	178.6711	3.275037	5084.600
A9V	4022	10.0574	0.223409	1686.467

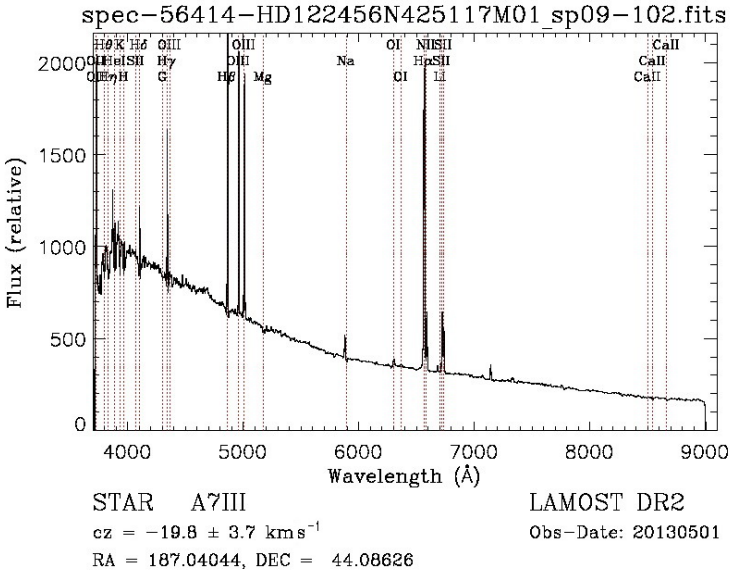

Fig. 4 An example of a spectrum with emission lines.

cal relative density of a data point d is the ratio of the average local reachable density of the neighboring point d to the local reachable density of data point d. Using the definition in Equation (2), we can calculate the distance of each model in D and the k-local outlier factor for each object d.

$$\mathrm{lrd}_k(d) = 1/\frac{\sum_{o\in N_k(d)} \mathrm{nearest}_k(d,o)}{\mid N_k(d) \mid}. \qquad (2)$$

The LOF algorithm needs to calculate the distance between two data points, resulting in a time complexity of $O(n^2)$ for the entire algorithm. The LOF value is calculated for each subset. LOF values with abnormal score less than or equal to a preset threshold are removed from the data set, and the rest are used to find a more suitable nearest neighbor, and then the LOF values are updated. For each data point, its distance from all other points is

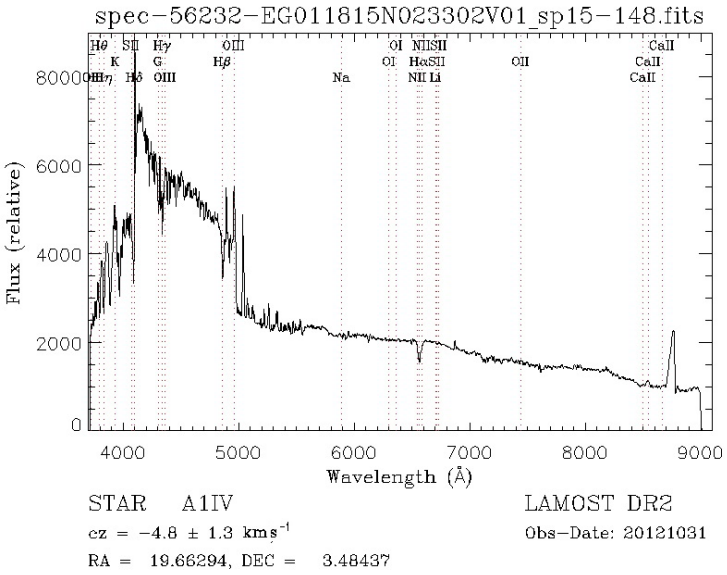

Fig. 5 An example of a spectrum with bad flux calibration.

calculated and the distances are sorted from near to far, as shown in Equation (2), its k-nearest-neighbors are located and finally LOF for the data are generated as shown in Equation (3).

$$\mathrm{LOF}_k(d) = \frac{\sum_{o \in N_k(d)} \frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}_k(d)}}{\mid N_k(d) \mid}. \tag{3}$$

4 OUTLIER ANALYSIS

Data mining is the process of acquiring knowledge from vast amounts of data. With the continuation of the LAMOST survey (the first year pilot survey and several years of general survey), astronomers are devoted to extracting valuable information from observational data to separate peculiar A-type stars from the normal identifications of objects. Classification of peculiar A-type stars mainly relies on the associated photometric and spectroscopic observations. Based on experiments that analyze subsets of data extracted from big data, and generate knowledge by analyzing the line indices of 144 340 A-type stellar spectra, including a total of 18 subclasses from A0 to A9, the distribution of Mahalanobis distance is calculated by considering the interdependencies of different subclasses. The results are consistent with what AstroStat provides via clustering analysis.

Table 2 displays example calculations of the main parameters associated with the outlier data including the Mahalanobis distances and spectral filenames.

The outliers can belong to small or sparse clusters, or might not belong to any clusters. We notice some outliers from Figure 3. After clustering the star data, we can consider whether the physical characteristics of the spectra in the cluster are obvious and consistent, and a mean value spectrum is introduced to help analyze the outlier data. There are two aspects involved in exploring the spectral information, one is the wavelength information, the other is the flux information. These two pieces of information correspond to the coordinates of the X-axis and the Y-axis, used in the process of plotting the function, and each wavelength value is associated with the flux value. To check for outliers, we read the spectra with R language and visually check them.

4.1 Spectral Data from an Emission Line Star

In the spectrum of an A-type star, there are strong emission lines, and their line indices are negative. It is easy to identify emission line stars as outliers. Generally, emission lines of a single star are produced by nearby thin gas, but these gases extend a very small range and the observer cannot separate them from the stars. Figure 4 is an example of an emission line star.

4.2 Broken-Spectrum Data

The process of acquiring data in the LAMOST survey causes instabilities to exist. Due to a split between spec-

Table 2 Outlier Spectral Data Based on Line Indices

Obsid	S/N	Subclass	kp12	$h_\delta 12$	$h_\gamma 12$	$h_\beta 12$	Mahalanobis distance	Spectral filename
106608008	12.31	A5V	6.79	12.84	17.64	36.07	5200.643808	spec-56306-GAC080N33M1_sp08-008.fits
100914163	341.02	A9	56.14	19.55	5.6	5.41	5084.600964	spec-56295-VB056N24V1_sp14-163.fits
105810249	11.67	A5	33.21	–17.09	–1.16	3.33	3920.594197	spec-56304-GAC094N27M1_sp10-249.fits
36710084	34.04	A2V	51.43	11.35	0.7	5.12	3149.108782	spec-55960-F5596001_sp10-084.fits
149409102	33.51	A7III	9.38	11.42	15.61	28.06	2560.394258	spec-56414-HD122456N425117M01_sp09-102.fits
72905248	10.92	A7IV	–9.7	1.55	3.57	5.58	349.672845	spec-56225-GAC100N32M1_sp05-248.fits
37110225	15.72	A1IV	–0.94	–1.18	1.96	3.85	317.2299728	spec-55960-GAC_101N09_V2_sp10-225.fits
15514122	10.48	A5	15.75	–0.51	2.9	5.28	249.4960056	spec-55910-GAC_078N28_B1_sp14-122.fits
75415148	35.71	A1IV	11.8	12.01	8.94	8.85	208.1938896	spec-56232-EG011815N023302V01_sp15-148.fits
15515102	15.59	A1IV	16.65	2.52	3.63	5.39	81.53144007	spec-55910-GAC_078N28_B1_sp15-102.fits

trometers in LAMOST, the two wavelength ranges can be used to locate errors in the spectrum. These data are often referred to as broken-spectrum data, which arise from a sudden loss of flow from the survey or instability in the data flow. Broken-spectrum data are caused by instability in the acquisition equipment or spectral errors, and are defined as "dirty data." We need to identify "dirty data" cases from survey data which introduce errors into a spectrum.

Although our mining works use the line index to avoid the effect of continuum fitting, which is a complex process in spectral analysis, some bad spectra with line indices may still be present. Figure 5 is an example of a bad spectrum that results from calculation of the line index.

5 SUMMARY

In this paper, k-means is employed for data mining spectra from the LAMOST survey by implementing clustering and outlier analysis. The AstroStat statistical tool is successfully applied to the LAMOST DR2 dataset and the Lick line index of the survey data is taken as the feature and clustered by the k-means algorithm. Mahalanobis distance analysis combines spectral data distance statistics and similarity measures. More than 140 000 spectra of A type stars are clustered to help identify spectra which do not follow the distribution of most A-type spectra. The line index plays an important role in the clustering process and can fully preserve the physical characteristics of the spectral data. AstroStat is an efficient tool for data mining and identifying bad data in order to separate rare and anomalous spectra from normal ones.

Acknowledgements We are very grateful to the anonymous referee for many useful comments and suggestions. This work is supported by the Joint Research Fund in Astronomy (U1631239) under cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS). It is also supported by the International Science and Technology Cooperation Program of China (2014DFE10030) and the Basic Science and Engineering Special Project of Heilongjiang Province Education Department (135109219).

References

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. 2000, in ACM DIGMOD International Conference, 29, 93

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1197

Ihaka, R., & Gentleman, R. 1996, Journal of computational and graphical statistics, 5, 299

Kembhavi, A. K., Mahabal, A. A., Kale, T., et al. 2015, Astronomy and Computing, 11, 126

Liu, C., Cui, W.-Y., Zhang, B., et al. 2015, RAA (Research in Astronomy and Astrophysics), 15, 1137

Luo, A.-L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1243

Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA (Research in Astronomy and Astrophysics), 15, 1095

Macqueen, J. 1967, Some Methods for Classification and Analysis of MultiVariate Observations[C], Proc. of, Berkeley Symposium on Mathematical Statistics and Probability, 281

Wei, P., Luo, A., Li, Y., et al. 2014, AJ, 147, 101