Clustering analysis of line indices for LAMOST spectra with AstroStat

Shu-Xin Chen¹,², Wei-Min Sun¹ and Qi Yan¹

¹ Key Lab of In-fiber Integrated Optics, Ministry Education of China, Harbin Engineering University, Heilongjiang Harbin 150009, China; sunweimin@hrbeu.edu.cn
² College of Mechanical and Electrical Engineering, Qiqihar University, Heilongjiang Qiqihar 161006, China

Received 2018 February 3; accepted 2018 April 13

Abstract The application of data mining in astronomical surveys, such as LAMOST (Large Sky Area Multi-Object Fiber Spectroscopy Telescope) survey, provides an effective approach to automate analyze the large amount of survey data with much complexity. Unsupervised clustering could help astronomers finding the association and outliers in a big data set. In this paper, we employ k-means method to perform clustering line index of LAMOST spectra with the powerful software AstroStat. The line index of the astronomical spectra is an effective way to extract spectral features for low resolution spectra, which can get the main spectral characteristic of stars. There are totally 144,340 line indices of A type stars are analyzed through calculating their intra and inter distances between pairs of stars. For intra distance, we use Mahalanobis definition to explore the cluster degree for each class, while for outlier detection, we define a local outlier factor (LOF) for each spectra. 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 few rare astronomical objects. We show two examples of these outliers, a spectra with abnormal continuum and a spectra with emission lines. Our work exhibit that line index clustering is a good method for data quality examination and rare object detection.

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

1 INTRODUCTION

Due to the limitation of observational equipment, the traditional astronomy and astrophysics research are basically based on small samples. With the improvement of telescope observational capabilities, 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 the survey data become bigger and bigger. A representative telescope is the LAMOST, which took the lead in the world for its efficiency to obtain celestial spectra (Cui 2012). The Galactic surveys of LAMSOT produce large amount of stellar spectra (Luo2012), (Luo2015), and there are more than 7 million spectra observed from Nov 2011 to June 2016, and the Fourth Data Release (DR4) has been released (http://dr4.lamost.org). LAMOST spectral analysis pipeline uses templates for stellar classification according to the similarity of observed with templates (Wei2014).

Template based classification is a supervised approach, and all data can be grouped to the class that template labeled, which would result in some rare objects dropping into presupposed class and could not
be detected. Data mining can automatically handle the data analysis for a large amount of data, revealing hidden, previously unknown and potentially valuable information (Liu2015). The application of data mining method, including clustering, outlier analysis and feature learning could be used to mining the data and discover new knowledge. A very useful data mining tool, AstroStat (KembhaviAK2015), has been developed since 2009, and is employed in this work to play with the LAMOST stellar spectra, which try to find the performance of the application of AstroStat in 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, and the concept of clustering was earliest defined. By using the process of aggregating similar data together, the unaggregated data are obviously different. The unsupervised clustering algorithm does not require the step of learning in advance, while the data need to be preprocessed. Due to the diversity of the astronomical data structure and feature, unsupervised clustering algorithm are suitable to recognize the inherent distribution of the data and the hidden knowledge pattern without providing the classification information. Analyzing the 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 works. As we know that absorption and emission lines are important features in a spectrum. The line index system is defined as a powerful feature extraction tool, and a serials of values of line indices represent a spectrum with specific physical characteristics. The line index is measured from the equivalent width (EW) 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. The widely used line index system is the Lick Line Index, which have been applied in dealing with many survey data such as SDSS. In this paper, the Lick line index is used in the LAMOST spectral data to extract atmospheric physical parameters, since the line index was almost unaffected by flux calibration errors and redshifts without the need for 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 pseudo-continuous spectrum. In LAMOST data release (Wei2014), the line indices used in our research are provided in the form of arrays on the LAMOST 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 follow. In section 2, we introduce the AstroStat and the line indices data of LAMOST. In section 3, we describe the k means method, intra-cluster correlation analysis with Mahalanobis distance, and using LOF to detect the outlier measurement with distance.In section 4, we search for outliers through the data mining process including using the 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 is based on Java, developed by the Virtual Observatory-India (VOI) project allowing astronomers to use both simple and sophisticated statistical routines on large datasets. It is an easy-to-use software tool for statistical analysis and visual description on big data. The users can freely download from http://vo.iucaa.ernet.in/voi/VOStat.html. AstroStat loads the VOPlot service and the current active data file, and use for any kinds of data visualization. Some points of interest can be noted directly in the plot, which provide the possibility to select these points for manual analysis. We use VOPlot in AstroStat to realize the data analysis with visualization tools which provided by LAMOST. AstroStat shows the toolbar and K-means partitioning primary interface in the top with which allows the user to select columns from multiple files in Fig.1.
Fig. 1 A screenshot of AstroStat showing the toolbar at the top and K-means Partitioning primary interface which allows the user to select columns from multiple files.

2.2 R language and LAMOST FITS

Data mining algorithms in AstroStat are based on R language. The R language came into existence as a free counterpart of the S statistical language from Bell Labs. Ross Ihaka and Robert Gentleman (R1996) 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 R language, and since R is a versatile open-source system for statistics so that integrates multiple data analysis and visualization methods. We decided to use R as the preprocessing step of LAMOST data. It not only has powerful data analysis capabilities, but also can effectively simplify the data analysis process.

The first step of preprocessing is reading the LAMOST FITS by using the RFITSIO package in R. Although AstroStat integrate the FITS reading package, we prefer to use the RFITSIO directly to avoid the problem of non-standard format of LAMOST spectra. The FITS data of LAMOST are written by the cfitsio package, and the FITS file name is "spec-MMMMM-YYYY_spXX-FFF.fits" (Luo2012). Downloading the data and loading the required FITSio package with the function require (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 Data columns, and limits the maximum data range value. Use the function plot () to select type = "s" and select the A0III star spectral data file spec-55976-GAC_099N04_V5_sp12-128.fits in the LAMOST survey database, as shown the plot of the flux spectrum in Fig. 2.
2.3 Data

LAMOST process the spectra to have the same starting and ending wavelength scale at 3800 to 9000Å. In the date release, the Lick line indices are provided for A type stars. We download 144,340 A type stars with line indices released. The important goal of astronomy is to discover anomalous, sparse and even unknown types, and as the important means, outlier data mining can effectively find out the feature anomaly and the relatively tiny differences from the surveyed day data. Spectral lines are important main features in spectra especially for A type stars because their continua is 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, clustering analysis is more suitable for data characteristics, and has been widely used in astronomical research. Qualitatively analysis the clustering result is a key step for correlation analysis and outlier detection. By use the merit of Mahalanobis distance
measurement, we don’t need to consider the scale difference of line indices and are easy to find those spectra with small difference in features. In addition, for obvious outliers in a high dimensional space, the LOF factor with whatever distance measure could be quantitatively used for the rare object detection.

3.1 K-means

K-means is one of the simple unsupervised learning algorithm that could solve the classification problems used by MacQueen (1967). The classify given data by finding the clusters and the centers. Then each set of line index is assigned to the nearest center that is closest in a least square sense. After all spectra are loaded, each point is replaced by the respective cluster center. Firstly, the algorithm initializes the clustering centers and normalizes the data. Theoretically, the initial clustering centers are found in the remaining data sets by using the maximum distance between each two pairs excluding those isolated points. Practically, the number of the 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 clustering centers are selected, multiple iterations are carried on, excluding the isolated points by checking if they exceed the threshold of a certain class. Finally, the K-means provides the clusters from all the spectra assigned to one of them. The threshold range should be set to keep most data inside the two classes. The clustering centers of the K-means algorithm guarantee that the objects in the same cluster are similar, while the objects in the different clusters are not similar.

The data in Fig.3 are from the LAMOST second data release (DR2) (http://dr2.lamost.org). Different line indices are clustered by K-means algorithm using AstroStat, and Fig.3 illustrates that most of A type stars distribute in a small local area in the plane of kp12 v.s. H−delta12. The clustering processing software platform in AstroStat is based on open source of efficient R language programming, and the Fig.3 shows the result of clustering in the line index plane Kp12 and H−delta.

3.2 Intra-cluster Correlation Analysis with Mahalanobis distance

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

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

The experiment uses Mahalanobis (data, center = Avg, cov = S) provided by R language, where Avg is the mean of center, and S is the cov sample covariance matrix. The function represents the Mahalanobis distance between each piece of data and the global library. The transformation of Mahalanobis distance is similar to that of principal component analysis (PCA) solution, ie, 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-negative, reflexivity, symmetry and trigonometric inequality. If the covariance matrix is a unit matrix, it is reduced to the Euclidean distance. We calculated Mahalanobis distance of each A-subtype for the dataset used in figure 3, and the distances of median, the maximum and the minimum of each subtype are shown in Table 1.
In the process of experimental data processing, the Mahalanobis distance need not be normalized, but the Euclidean distance must be normalized first and then calculate the distance between each other, otherwise the distance value is meaningless. Although the Euclidean distance is more commonly used, the flow values of each spectrum represent different characteristics. The Mahalanobis distance can better reflect the importance of different eigenvectors.

### 3.3 Outlier measurement with distance

Local Outlier Factor (LOF) is used for outlier detection (Markus2000), which can be applied to describe the singularity of un 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 p and an object o ∈ D such that Eq(2). LOF is according to 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 is subject to a specific probability distribution. However, the clustering method usually only gives the judgment of whether or not the abnormal point is "TRUE" or "FALSE", and can not quantify the degree of abnormality of each point. In contrast, density-based LOF algorithms are simpler and more intuitive. The analysis does not require too much data distribution to quantify outlierness. However, the LOF algorithm measures the anomalous degree of a data point not by looking at its absolute local density, but at the relative density of the data points with which it is nearby. The benefit which can data be distributed unevenly and with different densities. The local anomaly is defined by the relative local density. The local relative density of a data point d is the ratio of the average local reachable density of the neighbors of point d to the local reachable density of data.
Table 1 Analysis Mahalanobis distance of A type stars

<table>
<thead>
<tr>
<th>Subclass</th>
<th>Count</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td>608</td>
<td>22.08346</td>
<td>6.217810</td>
<td>627.5468</td>
</tr>
<tr>
<td>A0III</td>
<td>951</td>
<td>35.60493</td>
<td>1.879113</td>
<td>380.4221</td>
</tr>
<tr>
<td>A1IV</td>
<td>18312</td>
<td>14.56612</td>
<td>2.270596</td>
<td>1452.920</td>
</tr>
<tr>
<td>A1V</td>
<td>18060</td>
<td>15.48386</td>
<td>4.545271</td>
<td>966.1267</td>
</tr>
<tr>
<td>A2IV</td>
<td>14195</td>
<td>17.20922</td>
<td>0.215421</td>
<td>3149.108</td>
</tr>
<tr>
<td>A2V</td>
<td>2969</td>
<td>13.1970</td>
<td>0.839388</td>
<td>329.3868</td>
</tr>
<tr>
<td>A3IV</td>
<td>596</td>
<td>21.61400</td>
<td>1.061798</td>
<td>1073.728</td>
</tr>
<tr>
<td>A3V</td>
<td>16</td>
<td>960.2729</td>
<td>9.715015</td>
<td>10666.24</td>
</tr>
<tr>
<td>A5</td>
<td>24421</td>
<td>11.00235</td>
<td>0.696124</td>
<td>5200.643</td>
</tr>
<tr>
<td>A5V</td>
<td>14392</td>
<td>9.649854</td>
<td>0.226586</td>
<td>2560.394</td>
</tr>
<tr>
<td>A6IV</td>
<td>10970</td>
<td>14.91356</td>
<td>0.671972</td>
<td>706.7618</td>
</tr>
<tr>
<td>A6V</td>
<td>1709</td>
<td>11.08790</td>
<td>0.226586</td>
<td>2560.394</td>
</tr>
<tr>
<td>A7III</td>
<td>4873</td>
<td>9.291063</td>
<td>0.078180</td>
<td>894.8001</td>
</tr>
<tr>
<td>A7IV</td>
<td>25159</td>
<td>7.853723</td>
<td>0.098206</td>
<td>2087.251</td>
</tr>
<tr>
<td>A7V</td>
<td>1882</td>
<td>6.725759</td>
<td>0.39664</td>
<td>116.1464</td>
</tr>
<tr>
<td>A8III</td>
<td>33</td>
<td>178.6711</td>
<td>3.275037</td>
<td>5084.600</td>
</tr>
<tr>
<td>A9V</td>
<td>4022</td>
<td>10.0574</td>
<td>0.225409</td>
<td>1686.467</td>
</tr>
</tbody>
</table>

Using Eq(2) definition, we can calculate the distance of each model in D and the k-local outlier factor of each object d.

\[
lrd_k(d) = \frac{1}{\sum_{o \in N_k(d)} \text{nearest}_k(d, o)} \frac{1}{|N_k(d)|}
\]

LOF algorithm needs to calculate the distance between two data points, resulting in the entire algorithm time complexity of \(O(n^2)\). The LOF value is calculated for each subset. For those LOF abnormal score less than or equal to "True", removed from the data set, the rest to find a more suitable nearest-neighbor, and update the LOF value. First, for each data point, calculate its distance from all other points and sort it from near to far; as shown in Eq(2), find its k-nearest-neighbor and finally calculate the LOF data as shown in Eq(3).

\[
LOF_k(d) = \frac{\sum_{o \in N_k(d)} lrd_k(o)}{|N_k(d)|}
\]

4 OUTLIER ANALYSIS

Data Mining is the process of acquiring knowledge from vast amounts of data. With the ongoing of LAMOST survey (the first year pilot survey and three years of general survey), astronomers have devoted to extracting valuable information from the observational data to separate peculiar A−type stars from the normal identifications of objects. In the classification of peculiar A−type stars is mainly based on the photometric and spectroscopic observations. Based on experiments analyze data from big data, and mine knowledge of analyze the line indices of 144340 A−type stars spectrum, including a total of 18 subclasses from A0 to A9. The distribution of Mahalanobis distance calculated the interdependencies of different subclasses. The results were consistent with those of the outlier data, which AstroStat completed the clustering analysis. Table 2 is example calculate, which displays the main parameters of the outlier data including the Mahalanobis distance and spectral filename.

The outliers belong to small or sparse clusters, and without belonging to any clusters. We notice some outliers from Fig.3. After completing the clustering of the star data, it is analyzed whether the physical characteristics of the spectra in the cluster are obvious and consistent, and the mean value spectrum is introduced to help analyze the outlier data. There are two aspects of mapping the spectrum information, one is the wavelength information, the other is the traffic information. These two pieces of information correspond to the coordinates of the X-axis and the Y-axis, using in the process of drawing
the function, and each wavelength value corresponds of the flux value. To check the outliers, we use R language to read out spectra and visually check them.

### 4.1 Emission line star spectral data

In the A-type star spectrum, there are the strong emission line stars in the spectra, and their line indices are negative. It is easy to pick out 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 can not separate them from the stars. Fig.4 is an example of a emission line star.

### 4.2 Spectrum of abnormal continuous stars

During the shooting of the survey telescope, data exists instabilities. Due to split between spectrometers of LAMOST, the two wavelength ranges will find out the errors of spectrum. These data are often referred as the break-spectrum data, which generating a sudden loss of flow in the survey days or flowing instability. Broken-spectrum data is caused by instability of the shooting device or spectral errors, which is defined as "dirty data". In the survey data, we need to select those dirty data.

Although our mining works use line index to avoid effect of continuum fitting, which is a complex work in spectral analysis, some bad spectra with line indices may be better for false results calculation of line index.

### 5 SUMMARY

In this paper, the application of K-means on data mining in spectra of LAMOST survey is presented for the clustering and outlier analysis in massive survey data. 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 without dimension influence combines spectral data distance statistics and similarity measures, more than 140,000 spectra of the A type star are clustered to help to find out the spectrum which did not accord with the distribution of the most A star spectra. The line index plays an important role in the clustering process and can fully preserve the physical characteristics of the spectrum data. AstroStat is an efficient tool for data mining and bad data discovering to separate the rare and anomalous spectra from the 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); Basic Science and Engineering on special project of Heilongjiang Province Education Department(135109219)
Clustering Analysis of Line Indices for LAMOST Spectra with AstroStat

Fig. 4 An example of a spectrum with emission lines

References

Hou, W., Luo, A.L., Yang, H.F., et al. 2015, MN, 449, 276
Liu, C., Cui, W.Y., Zhang, B., et al. 2015, RAA, 15, 1139
Luo, A.L., Zhao, Y.H., Zhao, G., et al. 2015, RAA, 15, 1095
Markus, M., Breunig, H.K., Raymond, T.Ng., et al. 2000, ACM DIGMOD International Conference, 29, 93

This paper was prepared with the RAA L\LaTeX macro v1.2.
Fig. 5 An example of a spectrum with bad flux calibration