Cluster-Based Galaxy Spectral Analysis Post-print

Authors: Zhang Qian, Zhang Jiannan, Zhao Yongheng

Date: 2019-09-18T09:40:33+00:00

Abstract

Large-scale sky survey projects have generated massive astronomical datasets, necessitating research into automated spectroscopic processing methods suitable for large-scale data. Traditional galaxy spectral classification methods based on line detection or BPT diagrams are difficult to directly apply to automated galaxy spectral classification pipelines; by contrast, machine learning-based automated spectral analysis is more suitable for classification research on massive astronomical datasets. This paper proposes a galaxy spectral analysis method based on two-layer clustering. The first layer employs the k-means clustering algorithm to cluster galaxy spectra into absorption-line galaxies and emission-line galaxies, while the second layer uses the CLARA (Clustering LARge Applications) clustering algorithm to cluster emission-line galaxies into five clusters. Experiments conducted on galaxy data from LAMOST DR5 demonstrate that: (1) the first-layer k-means clustering can successfully separate galaxy spectra into absorption-line galaxies and emission-line galaxies, with clustering results being essentially consistent with classification based on line detection; (2) the second-layer CLARA clustering results can reflect different galaxy types in BPT diagrams; (3) the spectral clustering results exhibit the expected correlation with color-magnitude diagram classification; and (4) both k-means clustering and CLARA clustering are applicable to automated analysis and processing of large-scale data, with clustering results effectively reflecting the physical properties and evolutionary processes of galaxies, and cluster center data can provide templates for automated spectral classification pipelines.

Full Text

Spectral Classification of Galaxies Based on Clustering Analysis

Xi Zhang[12], Jiannan Zhang[1], Yongheng Zhao[1]

[1]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Large-scale sky survey projects have produced massive astronomical datasets, necessitating the development of automated spectral processing methods suitable for big data. Traditional galaxy spectral classification methods based on line detection or BPT diagrams are difficult to apply directly to automated galaxy spectral classification pipelines. In contrast, machine learning-based automated spectral analysis is more suitable for classification studies of massive astronomical data. This paper proposes a galaxy spectral analysis method based on double-layer clustering. The first layer employs the k-means clustering algorithm to separate galaxy spectra into absorption-line galaxies and emission-line galaxies, while the second layer uses the CLARA (Clustering Large Applications) algorithm to further cluster emission-line galaxies into five subtypes. Experiments conducted on galaxy data from LAMOST DR5 demonstrate that: (1) The first-layer k-means clustering successfully separates galaxy spectra into absorption-line and emission-line galaxies, with results largely consistent with classification based on spectral line detection. (2) The second-layer CLARA clustering results reflect different galaxy types in BPT diagrams. (3) The spectral clustering results show expected correlations with color-magnitude diagram classifications. (4) Both k-means and CLARA clustering are applicable to large-scale automated data analysis, with clustering results effectively reflecting the physical properties and evolutionary processes of galaxies, and cluster centers providing templates for automated spectral classification pipelines.

Keywords: LAMOST; Clustering; Galaxy spectral classification; Large-scale spectral analysis

---

Galaxy spectral classification is crucial for studying galaxy formation and evolution. Traditional classification methods include: Hubble' s morphological classification, which categorizes galaxies into elliptical, spiral, barred spiral, and irregular types based on appearance; color-based classification, where Strateva et al. [**?**] found that color-magnitude diagrams from SDSS data exhibit a bimodal distribution with distinct peaks for blue and red galaxies, separated by a "green valley" ; and spectral classification using the Baldwin, Phillips, & Terlevich (BPT) diagnostic diagram [**?**], which has evolved into line-ratio diagnostic diagrams with empirical demarcation lines such as Kauffmann' s line for identifying pure star-forming galaxies [**?**], Kewley et al.' s line for identifying pure AGN galaxies [**?**], and lines by Kewley [**?**] and Cid Fernandes [**?**] for distinguishing LINER (Low-Ionization Nuclear Emission-Line Region) galaxies from Seyfert 2 galaxies.

Large-scale survey projects have provided massive spectral datasets for astronomy, including 2dF, 6dF, RAVE, SDSS, LAMOST, and GAIA. LAMOST DR5 alone released over 150,000 galaxy spectra, necessitating research into auto-

mated spectral classification techniques for large-scale data processing. Traditional methods based on line detection or BPT diagrams require stellar population synthesis, a complex and time-consuming process unsuitable for massive datasets and incompatible with automated classification pipelines. In contrast, machine learning-based spectral classification methods are better suited for analyzing massive astronomical data. Numerous machine learning approaches have been successfully applied to astronomical classification, including both supervised and unsupervised methods. Unsupervised methods include Principal Component Analysis (PCA), widely used for galaxy spectral identification and classification—for example, the spectral processing system in the SLOAN survey identifies galaxy spectra using principal components [**?**], and Almeida et al. [**?**] successfully applied k-means clustering to galaxy spectral classification, with results effectively reflecting galaxy evolutionary processes. Supervised methods include Fisher discriminant analysis for quasar and normal galaxy classification [**?**], Support Vector Machines for active and non-active celestial object classification [**?**], and decision trees for galaxy morphological classification [**?**].

Clustering, as an unsupervised method, offers simplicity, fast convergence, and high accuracy. It relies primarily on data features for automatic classification, operating independently with minimal subjective influence. Unlike supervised methods, it does not require labeled training data, and clusters with smaller populations can help identify rare celestial objects. This paper designs a double-layer clustering method for analyzing galaxy spectral data from LAMOST DR5. The structure is as follows: Section 1 introduces the double-layer clustering method, Section 2 describes the galaxy spectral clustering experiments including preprocessing, procedures, and parameter selection, Section 3 analyzes the experimental results by evaluating effectiveness and physical significance through comparisons with classifications based on line detection, BPT diagrams, and color-magnitude diagrams, and Section 4 presents the conclusions.

1. Double-Layer Clustering Method

Based on the characteristics of galaxy spectra and different clustering algorithms, this paper proposes a double-layer clustering method for galaxy spectral analysis. The first layer uses the k-means clustering algorithm [**?**] to separate galaxy spectra into absorption-line and emission-line galaxies. K-means is simple, converges quickly, and scales well for large datasets, making it suitable for massive galaxy spectral processing. The second layer employs the CLARA clustering algorithm [**?**] to further divide emission-line galaxies into five subtypes. CLARA is simple, robust to noise, and suitable for large-scale data.

1.1 K-Means Clustering Algorithm

The core of the k-means clustering algorithm is to partition n samples into k clusters while minimizing the sum of squared distances from each sample point to its cluster center. The basic steps are:

Input: n samples and number of clusters k.
Output: Partition of samples into k clusters.

1. Select k initial points from n samples as initial cluster centers.
2. Calculate the distance between each sample point and all cluster centers, assigning each sample to the nearest cluster center.
3. Recalculate the mean of all samples in each cluster as the new cluster center, and compute the sum of squared distances D from each sample to its cluster center.
4. Check if cluster centers or D have changed. If changed, update cluster centers and repeat steps 2-3; otherwise, clustering ends.

Many factors affect clustering performance, including k value selection, initial center selection, and distance metric. K value can be chosen empirically or based on density. Common methods for selecting initial cluster centers include: (1) random selection of k samples; (2) random selection of 10% of data for pre-clustering with random initial centers; (3) uniform random selection of k centers within the sample space range; and (4) weighted k-means++ method, where after randomly selecting the first center, distances from all points serve as weights for selecting subsequent centers, giving higher probability to distant points. Distance metrics include Euclidean, Manhattan, cosine, and correlation distances.

In our experiments, considering spectral characteristics and comparing multiple distance measures, we selected correlation distance as the metric. The correlation distance is defined as $d = 1 - \rho$, where ρ is the correlation coefficient measuring the relationship between random variables X and Y, with range [-1,1]—larger absolute values indicate stronger correlation.

1.2 CLARA Clustering Algorithm

K-means is sensitive to noise. The k-medoids algorithm [**?**] improves upon k-means by replacing cluster centers with actual data points rather than means, reducing outlier impact. The basic steps are:

Input: n samples and number of clusters k.
Output: Partition of samples into k clusters.

1. Select k initial points from n samples as initial cluster centers.
2. Calculate distances from all samples to cluster centers, assigning each to the nearest center.
3. Randomly select a non-center point, calculate the total cost of replacing the original center with this point, and repeat until all non-center points are evaluated.
4. If any replacement yields negative total cost, select the one with minimum cost as the new center.
5. Repeat steps 3-4 until cluster centers stabilize.

To determine whether a new non-center point O_h can replace original center O_i,

each non-center point O_j must satisfy: after replacement, O_j is assigned to the nearest cluster (which could be O_i, O_m, or the new O_h). The total replacement cost is the sum of costs for all non-center objects:

$$C_{Tih} = \sum_{j=1}^{n} C_{jih}$$

where C_{jih} is the cost for O_j when O_i is replaced by O_h, defined as the difference between distances to the original and new centers. Negative total cost indicates replacement is beneficial; positive cost means no change is needed.

Since k-medoids requires exhaustive search for optimal solutions, it only works for small datasets. CLARA (Clustering Large Applications) improves k-medoids by using sampled data to represent the full dataset for center calculation, enabling large-scale clustering.

CLARA Algorithm Steps:
Input: n samples, number of clusters k, number of samples m.
Output: Partition of samples into k clusters.

1. Repeat m times: randomly sample $(40 + 2k)$ points from the full dataset and perform steps 2-4.
2. Apply k-medoids to the sample to select k cluster centers.
3. Calculate distances from all non-center points to these centers and assign them to nearest clusters.
4. Compute total cost from step 3. If cost is lower than current best, apply these centers to the full dataset; otherwise return to step 1 for the next iteration.

2. Galaxy Spectral Clustering Experiment

2.1 Data Preprocessing

This study uses 30,000 randomly selected spectra from 153,093 galaxy spectra in LAMOST DR5.

Due to the lack of photometric calibration equipment, LAMOST employs relative flux calibration using high-quality F-type dwarf stars as standards to derive instrument response curves. However, reddening of these standard stars can cause continuum uncertainties, requiring spectral recalibration. We use SLOAN u,g,r,i,z fiber magnitudes to recalibrate LAMOST continua.

After recalibration, spectra are deredshifted to rest wavelengths and resampled over 3600-9000Å with 1Å intervals. To avoid noise and environmental effects, flux normalization is applied. Assuming a spectrum is an n-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_n)$, the normalization method [**?**] is:

$$y_i = \frac{x_i}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

After removing spectra that cannot be recalibrated or have invalid redshift values, 27,272 galaxy spectra remain for clustering experiments.

2.2 Clustering Experiments

K-means and CLARA clustering algorithms are applied to LAMOST DR5 galaxy spectra in two layers. The first layer uses k-means to separate spectra into absorption-line and emission-line galaxies. The second layer uses CLARA for fine classification of emission-line galaxies.

First Layer: K-means clustering partitions the 27,272 preprocessed spectra into emission-line and absorption-line galaxies. Early-type galaxies dominated by old stars show strong absorption lines with weak or undetectable emission lines, while some late-type galaxies exhibit similar weak emission features. In later-type galaxies, absorption lines gradually lose dominance as emission lines become more prominent. To emphasize emission and absorption features, continua are removed using median filtering to fit and subtract the continuum, leaving line information for clustering. Considering galaxies with both emission lines and stellar components, we set k=3, use k-means++ for initial center selection, and employ correlation distance.

Second Layer: CLARA clustering further subdivides emission-line galaxies from the first layer. Since continua reflect some emission-line galaxy characteristics, they are retained. We select 12,689 spectra with r-band S/N > 5. To avoid sky line effects, median filtering with a 5-point window is applied. As some samples only have flux values in 3600-7900Å and CLARA depends on sample points, we use this wavelength range. With 100 sampling iterations and correlation distance, we determine the optimal k using the elbow method on the SSE (sum of squared errors) curve, observing a clear elbow at k=5.

[Figure 1: see original paper] shows the SSE variation with k value.

3. Analysis of Galaxy Spectral Clustering Results

3.1 First-Layer Clustering Results Analysis

K-means clusters the 27,272 spectra into three clusters (cluster1, cluster2, cluster3). Their cluster centers [Figure 2: see original paper] reveal galaxy types: emission-line galaxies are dominated by emission lines, with cluster1 showing strong emission lines from galaxies with weak stellar components; absorption-line galaxies show dominant absorption lines with weak or undetectable emission lines, identifying cluster2 as absorption-line galaxies; cluster3 shows weak emission lines from galaxies with stellar components.

To assess stability, k-means is applied to four S/N subsets: r-band S/N > 5, 10, 15, and 20, containing 23,465, 15,593, 9,120, and 5,166 spectra respectively. The resulting cluster centers [Figure 3: see original paper] consistently identify emission-line, absorption-line, and weak emission-line galaxies across all subsets, demonstrating k-means stability.

[Figure 4: see original paper] shows distance distributions between each spectrum and cluster centers. Overall, cluster i is closest to its own center. The left column shows cluster1-center1 distances peak near 0, with distant peaks for other centers, indicating clear separation. Clusters 2 and 3 show closer distances to their own centers within each S/N subset. As S/N increases, distances to their own centers approach 0 (e.g., cluster2-center2 peak decreases from 0.65 to 0.4), though cluster2 and cluster3 show less compact distributions than cluster1.

Comparing with traditional classification: conventional emission/absorption-line discrimination uses line S/N ratios. References [**?**, **?**] filter on Hα, Hβ, [OIII]$ 5007, $and[NII]$ 6585, $butCidFernandesetal.[**?**]foundthisexcludesweakemission-linegalaxies.Therefore, weonlyfilteronH $ with S/N > 3. In our results, cluster1 and cluster3 are emission-line galaxies, cluster2 is absorption-line. Comparison shows agreement rates of 97.79%, 80.80%, and 84.52% per cluster, with 89.0% overall agreement.

Color-magnitude diagrams for each cluster [Figure 5: see original paper] show the bimodal distribution with red and blue peaks separated by a green valley. Emission-line cluster1 occupies the blue region, absorption-line cluster2 the red region, and weak emission-line cluster3 the green valley, consistent with early-type galaxies being predominantly red and late-type galaxies blue.

These results demonstrate that k-means efficiently classifies galaxy spectra into absorption-line and emission-line types, converging quickly even for large datasets. The clustering reflects physical properties and aligns with traditional methods, proving its feasibility. Cluster centers provide templates for automated classification pipelines with stronger noise resistance than line-analysis-based templates.

3.2 Second-Layer Clustering Results Analysis

CLARA subdivides emission-line galaxies into five subtypes (emi1-emi5) , with cluster centers being actual spectra from each class [FIGURE:6, first column]. Distance statistics show each cluster is closest to its own center (peaking near 0) with five distinct peaks, indicating clear inter-class separation.

Comparison with BPT diagram classification: BPT methods measure line strengths of Hα, Hβ, [OIII]$ 5007, $and[NII]$ 6585.$Galaxyspectraareconsideredcombinationsofstellarspectra.
$frequencybackgroundwith$201-pointmedianfiltering, andfitsingleGaussianstoH $, [OIII]$ 5007, $and[NII]$ 65
Line intensity is calculated using:

$$I = \int_{\lambda_1}^{\lambda_2} (F(\lambda) - C(\lambda))d\lambda$$

where $F(\lambda)$ is observed flux and $C(\lambda)$ is continuum. Due to high spectral quality requirements and weak emission lines, only 8,122 emission-line galaxies could be BPT-classified. Results for emi1-emi5 are shown in and [FIGURE:6, middle column], with background density showing all emission-line galaxies and red points showing each class.

The BPT diagram includes: Kauffmann' s K03 line [**?**] (Equation 5) for pure star-forming galaxies (below line); Kewley' s K01 line [**?**] (Equation 6) for pure AGN (above line); and Cid Fernandes' CF10 line [**?**] (Equation 7) separating Seyfert 2 (above) from LINER (below) galaxies.

$$\log([\text{OIII}]/\text{H}\beta) = 0.61/(\log([\text{NII}]/\text{H}\alpha) - 0.05) + 1.3 \quad \text{(K03)}$$

$$\log([\text{OIII}]/\text{H}\beta) = 0.61/(\log([\text{NII}]/\text{H}\alpha) - 0.47) + 1.19 \quad \text{(K01)}$$

$$\log([\text{OIII}]/\text{H}\beta) = 0.01 \times \log([\text{NII}]/\text{H}\alpha) + 0.48 \quad \text{(CF10)}$$

Emi1 distributes mostly below K01, including star-forming and composite galaxies. Emi2 is predominantly below K03 (84.00% star-forming). Emi3 resembles emi1 but contains more AGN. Emi4 is below K03 (84.31% star-forming) but with larger [OIII]/Hβ ratios, showing stronger emission lines, flatter continua, and weaker absorption than emi2. Emi5 contains 61.42% composite and AGN galaxies, with cluster centers dominated by stellar components and weak emission lines, contrasting emi2 and emi4' s emission-dominated centers. Overall, weaker stellar components correlate with stronger emission lines and star-forming region characteristics, evident in emi2 and emi4 centers.

Color-magnitude diagrams [FIGURE:6, right column] show star-forming galaxies (emi2, emi4) are bluer with stronger emission lines, emi1 and emi3 occupy the green valley, and emi5 leans red, consistent with AGN galaxies being predominantly early-type [**?**]. The progression from emi2/emi4 to emi1/emi3 to emi5 shows increasing AGN fraction and color transition from blue to red, supporting Schawinski' s [**?**] proposal that AGN activity suppresses star formation, potentially driving galaxies across the green valley.

BPT classification is complex and demanding, only classifying a subset of emission-line galaxies, while CLARA successfully partitions all spectra. CLARA' s advantages include low spectral quality requirements, no stellar component fitting, simplicity, and effectiveness for large-scale automated analysis, with results reflecting galaxy evolution.

4. Conclusion

For LAMOST DR5 galaxy spectra, k-means clustering successfully separates absorption-line and emission-line galaxies, consistent with line-detection-based classification. K-means is simple, efficient, and suitable for large-scale automated analysis, with results reflecting galaxy properties and providing noise-resistant templates for classification pipelines.

CLARA clustering further subdivides emission-line galaxies, showing expected correlations with BPT and color-magnitude classifications and reflecting evolutionary processes. CLARA requires low spectral quality, avoids stellar component fitting, and enables fast, effective automated classification of large datasets, providing templates for spectral classification pipelines.

Acknowledgments

The Guoshoujing Telescope (Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST) is a major national scientific project built by the Chinese Academy of Sciences and funded by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences. We thank the referee for questions and comments that led to in-depth discussions and improved manuscript rigor.

References

[1] Strateva I, Ivezic Z, G. R. Knapp, et al. Color separation of galaxy types in the sloan digital sky survey imaging data [J]. *The Astronomical Journal*, 2001, 122: 1861-1874

[2] J. A. Baldwin, M. M. Phillips, R. Terjevich. Classification parameters for the emission-line spectra of extragalactic object [J]. *Publications of The Astronomical Society of The Pacific*, 1981, 93: 5-19

[3] Kauffmann G, Heckman T M, White D M, et al. Stellar masses and star formation histories for 10^5 galaxies form the Sloan Digital Sky Survey [J]. *Mon. Not. R. Astron. Soc.*, 2003, 341: 33-53

[4] Kewley L J, Dopita M A, Sutherland R S, et al. Theoretical modeling of starburst galaxies [J]. *The Astronomical Journal*, 2001, 556: 121-140

[5] Kewley L J, Groves B. Kauffmann G. The host galaxies and classification of active galactic nuclei [J]. *Mon. Not. R. Astron. Soc.*, 2006, 372: 961-976

[6] Fernandes R C, Stasinska G, Schlickmann M S, et al. Alternative diagnostic and the 'forgotten' population of weak line galaxies in the SDSS [J]. *Mon. Not. R. Astron. Soc.*, 2010, 403: 1036-1053

[7] Bolton A S, Schlegel D J, Aubourg E, et al. Spectral classification and redshift measurement for the SDSS-III baryon oscillation spectroscopic survey [J]. *The Astronomical Journal*, 2012, 144: 144-164

[8] Almeida J S, Aguerri J A L, Munoz-Tunon C, et al. Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra [J]. *The Astronomical Journal*, 2010, 714: 478-504

[9] 李乡儒, 胡占义, 赵永恒. 基于 Fisher 判别分析的有监督特征提取和星系光谱分类 [J]. 光谱学与光谱分析, 2007, 27(9): 1891-1901

[10] 覃冬梅, 胡占义, 赵永恒. 基于支撑向量机的天体光谱自动分类方法 [J]. 光谱学与光谱分析, 2004, 24(4): 507-511

[11] Gauci A, Adami K Z, Abela J. Machine learning for galaxy morphology classification [J]. *Mon. Not. R. Astron. Soc.*, 2010, 000: 1-8

[12] Mac J. Some methods for classification and analysis of multivariate observations [C]. *Proceedings of 5th Berkeley Symposium on Mathematics, Statistics and Probability*, Berkeley, 1997: 281-296

[13] 赵国富, 曲国庆. 聚类分析中 CLARA 算法的分析与实现 [J]. 山东理工大学学报 (自然科学版), 2006(02): 45-48

[14] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values [J]. *Data Mining and Knowledge Discovery*, 1998, 2(3): 283-304

Note: Figure translations are in progress. See original paper for figures.