# A Performance Comparison of Unsupervised Clustering Techniques for Classification of Spitzer Space Telescope Infrared Spectra

*Bo Mu, Joel H. Kastner, and Catherine L. Buchanan*
*Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology*
*ROCHESTER, NY USA*

## Abstract

*Two unsupervised clustering techniques, hierarchical clustering and K-means clustering, have been investigated and their performances on classifying the Spitzer Space Telescope infrared spectra of stars have been compared. In order to reduce the data dimensionality without losing much information, Principle Components Analysis was applied to the scaled spectral data prior to the classification. The influences of different scaling methods have been evaluated as well. The least classification error is achieved by the hierarchical clustering technique with the average linkage applied to the data set, in which each spectrum is scaled by its maximum amplitude.*

## 1. Introduction

After their core hydrogen is exhausted via nuclear fusion, sun-like stars become very luminous red giants – thousands of times brighter than the Sun -- yet may be obscured visually by dusty, expanding circumstellar envelopes. The stars' photospheric emission is absorbed by these optically thick and dusty envelopes and re-radiated in the mid- and far-infrared range [1]. Hence, telescopes equipped with infrared sensors are an effective means to study the chemical enrichment of the Milky Way galaxy by the dusty envelopes of these mass-losing asymptotic giant branch (AGB) stars.

The Spitzer Space Telescope Infrared Spectrograph (IRS) is now providing high-quality infrared spectra with which we can investigate mass-losing AGB stars, especially in nearby, external galaxies such as the Large Magellanic Cloud (LMC). The LMC is an attractive region for study by astronomers because (1) it is our nearest neighbor galaxy, only 179,000 light years away; (2) its low metallicities and high star formation rates mimic those of far more distant, high-redshift galaxies and (3) it contains a large population of IR-luminous, mass-losing objects found at essentially the same distance, thereby alleviating the distance ambiguities that plague studies of mass-losing stars in the solar neighborhood. Therefore the IRS spectra of mass-losing AGB stars collected from the LMC can be used to establish broad-band photometric indicators of the envelope chemistry of mass-losing stars [2].

In the absence of detailed physical knowledge concerning the observed sources, unsupervised spectral clustering provides a path to identify the empirical similarities or dissimilarities among the sources so that we can begin to group them observationally. Unlike supervised clustering based on a training set of labeled data, unsupervised clustering seeks natural groupings in the data set without predefined target information, except for the number of desired classes. The clustering results are thereby completely based on the similarities of the observed data. Although many unsupervised clustering algorithms have been proposed, most of them are based on two popular techniques, K-means clustering and hierarchical clustering. K-means clustering is an iterative approach to find clusters and their centers such that the within-cluster sums of squared distance are minimized.

Hierarchical clustering is the other popular unsupervised classification technique. It expresses the data structure using a tree-shape diagram, or dendrogram. There are two basic approaches to the implementation of hierarchical clustering, agglomerative or divisive. The agglomerative approach sequentially merges individuals into groups, while the divisive approach sequentially separates individuals into finer groupings [3]. The agglomerative approach has been carried out to classify the LMC spectral data set because it is more easily implemented in practice.

To reduce the computational complexity and simplify the visualization of the data structure, a multivariate technique, Principal Component Analysis (PCA), can be applied to a high dimensional spectral data set. PCA transforms a number of related variables to a set of uncorrelated variables by applying the Single Value Deposition (SVD) technique to the covariance matrix of the data set or other similar techniques. The patterns in the data then can be found and the dimensionalities of the data can be reduced by means of mapping the high dimensional data into a lower dimensional uncorrelated vector space [4]. In the case of spectral classification of astronomical objects, given N variables (dimensions) for each spectrum of a set of objects, M (M <= N) new uncorrelated vectors (dimensions) can be constructed via PCA, such that each of their corresponding eigenvalues accounts for as much of the variance of the entire data set as possible. Then the projection of the spectral data set into the new uncorrelated vector space yields the underlying patterns of spectra.

Spectral classification algorithms are affected by factors such as how the spectral data set is scaled and how the similarity between clusters is measured. In this paper, we applied the above two unsupervised clustering algorithms with different combinations of factors to the LMC Spitzer spectral data set and evaluated their performances by comparing the classification results with the ground truth, which has been determined by Buchanan [2] and is available at http://www.cis.rit.edu/~clbsps/. Congalton [5] stressed that the Confusion Matrix, in which each row corresponds to the instances in actual classes while each column corresponds to the instances in theoretical classes, is a useful tool to assess the classifiers performance. Through the classifiers assessment, an optimum unsupervised clustering

approach for the LMC spectral data set can be found and applied to additional Spitzer space telescope infrared spectral data.

We describe the LMC spectral data set and the clustering procedure in section 2; section 3 is devoted to the performance comparison of the unsupervised classification algorithms and the discussion of classification results. The final section contains a summary.

## 2. Data and Experiment

A raw spectral data set, including 60 infrared spectra of the IR-luminous stars in LMC region, was obtained by the Spitzer IRS and processed using the Spitzer pipeline version S11.0 (see [2] for details). The clean spectral data set then was extracted by removing the sky background from the source spectra and made available as the Spitzer IRS Spectral Atlas of Luminous IR Sources in the LMC [2]. We carefully examined the spectral data set and selected 53 unambiguous spectra from the raw data set as our samples. In the selected spectral data set, two artifacts were corrected to facilitate further classification. The first artifact is that each spectrum has a different wavelength range. The other is the presence of overlapping sample values from 7.5 μm to 9 μm and from 20 μm to 22 μm. Thus, with the exception of two objects lacking short wavelength (5-14 μm) data (MSXLMC1072 and MSXLMC1524) all spectra were resampled onto the identical wavelength range, from 5.3382 μm to 33.097 μm with a single flux density value sampling at each wavelength interval. Although these two outliers cannot be classified, we still put them into an "Oddball" class (see below) to make the classification results easy to be compared with the ground truth. Thus the final data set consists of 53 samples including MSXLMC1072 and MSXLMC1524. It should be noted that all the source names in our data set come from Infrared Astronomical Satellite (IRAS) and Midcourse Space Experiment (MSX) catalogues.

Because different potential spectral scaling methods will undoubtedly have influence on our clustering results, it is essential to investigate them -- although consideration of the scaling method is often neglected in the classifier design. Three standard scaling methods were employed in our experiments: 1) scale each spectrum according to its mean and standard deviation to produce a new spectrum with zero mean; 2) scale each spectrum by its maximum amplitude, to compress its amplitude into the range [0 1]; 3) scale each spectrum by its area, so that all scaled spectral areas are identical. The scaled spectral data set and the original data set are plotted in Fig.1.

We can discern the typical features from the scaled spectra in Fig.2. These features reveal the chemistry of the circumstellar dust of the mass-losing evolved stars [2]. Spectra with a broad SiC dust emission peak at 11.5 μm and narrow acetylene ($C_2H_2$) absorption peak at 13.7 μm are characteristic of stars surrounded by Carbon-rich ejecta. Spectra with silicate dust emission peaks at 9.7 μm and 18 μm indicate that the stars have Oxygen-rich ejecta. Spectra with very red continua and narrow emission lines imply that the corresponding stars are young, luminous, and embedded in star-forming clouds. These objects are designated in [2] as "Red Objects" based on infrared color-color diagrams. The other sources, such as MSXLMC1072, MSXLMC1524, MSXLMC890, MSXLMC1326 and IRAS04553-6825 were put into an additional class ("Oddballs"): MSXLMC1072 and MSXLMC1524 have such a short wavelength range that they can not be classified; MSXLMC890 and MSXLMC1326 are B[e] hypergiants [6], which are young, massive stars and not red giants; IRAS04553-6825 is probably a highly obscured and luminous supergiant.

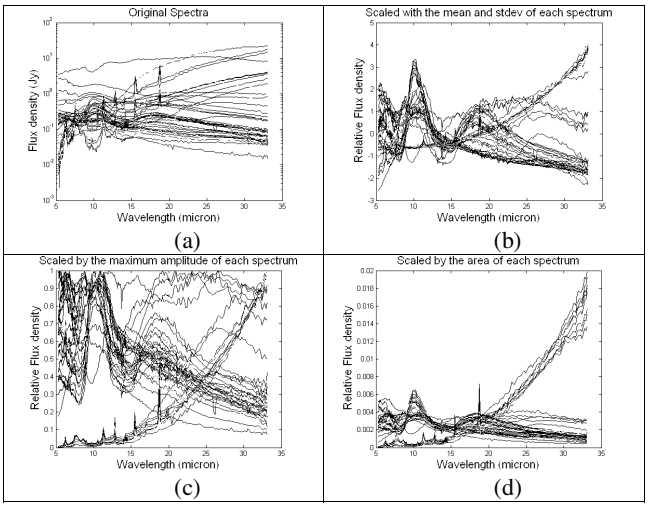


***Figure 1***. *Spectra data, (a). original, (b). each of them is scaled according to its mean and standard deviation, (c). each of them is scaled by its maximum amplitude, (d). each of them is scaled by its area.*

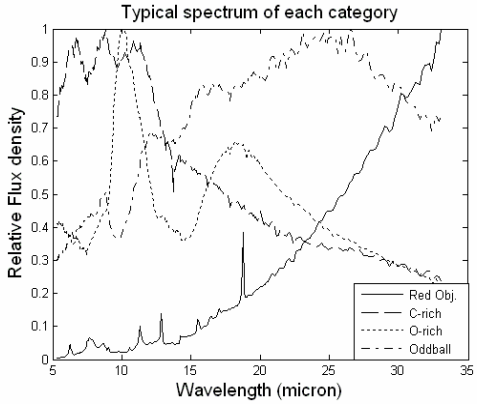


***Figure 2.*** *The typical spectrum of O-rich stars (MSXLMC141), C-rich stars (MSXLMC87), Red Objects (MSXLMC22) and "Oddballs" (IRAS04553-6825). Each spectrum is scaled by its maximum amplitude.*

The PCA method was then applied to the scaled spectra to reduce their dimensionalities. In this case, the first three Principal Components (PC) can express 95%, 97%, and 98% of the total variance using spectral scaling methods 1, 2, and 3, respectively. So the scaled spectral data sets can be projected onto the 3D space formed by the first three PCs without losing much information. This reduction in dimensionality results in a large gain in classification computational efficiency, which would be a great benefit in the case of a very large data set.

To use agglomerative hierarchical classification, we first need to find the similarities (Euclidean distance) between every pair of spectra in the data set. Then the spectra can be grouped into a hierarchical tree based on the proximity (Euclidean distance) between every pairs of objects. Single linkage, complete linkage and average linkage are three standard grouping methods.

Complete linkage, also called furthest neighbor, uses the largest distance between two clusters while average linkage uses the average distance [3]. These two linkage methods were employed in our experiment. Finally, the clusters can be chosen at a given hierarchical level. According to the classification results in [2], which are based on color-color diagrams, the recommended number of classes should be four and the clusters should be found at that hierarchical level.

The performance of another unsupervised clustering technique, K-means, has been investigated as well. Unlike hierarchical clustering, K-means clustering requires the initial cluster centers and the number of classes as the input parameters. The number of classes was fixed at four and the start points we chose are MSXLMC22, MSXLMC87, MSXLMC775 and IRAS04553-6825, which represent the typical features of each of the four clusters. Then the samples were labeled according to the nearest cluster centers. We iteratively computed the new centers and classified the samples until the centers could not be updated. At this point we have the final clusters and their respective centers.

**Table 1. The confusion matrix of hierarchical clustering using average linkage with PCA on the scaled spectra by the maximum amplitude**

| | | Ground Truth | | | | |
|---|---|---|---|---|---|---|
| | | C-rich | O-rich | Red | Odd | Total |
| Actual clusters | C-rich | 14 | 6 | 0 | 0 | 20 |
| | O-rich | 0 | 16 | 0 | 1 | 17 |
| | Red | 0 | 0 | 11 | 0 | 11 |
| | Odd | 1 | 0 | 0 | 4 | 5 |
| | Total | 15 | 22 | 11 | 5 | 53 |

***Note**: Red refers to Red Objects and Odd refers to Oddballs.*

The assessment of each clustering approach has been performed by comparing the classification results with the ground truth. The comparison results were tabulated in the confusion matrix; for example, Table 1 is the confusion matrix of the hierarchical clustering method using average linkage with PCA on the data scaled by the maximum amplitude. The overall classification accuracy can be approximated by the confusion matrix trace divided by the total number of spectral samples. The overall classification error is roughly the complement of the overall classification accuracy. For example, the overall classification accuracy reflected in Table 1 is about 84.9% and the overall classification error is about 15.1%.

## 3. Results and Discussion

As seen in Table 1, the hierarchical clustering algorithm using average linkage partitioned the Red Objects from the others without errors because their spectral patterns are very different from the other categories. If we interpret these differences in the color sense, the Red Objects are the most "reddish" (emission at long IR wavelength) while the C-rich stars are the most bluish (emission at short IR wavelength). The infrared 'color' of the O-rich stars resembles cyan, because their spectra have two peaks at short and middle IRS wavelengths, respectively. The stars belonging to the Oddball class have an emission peak at ~11.5 μm (in the short IR wavelength region) and a flat response over the long IR wavelength in their spectra. Thus the Red Objects are independent of the other three categories and can be easily separated from the others. Conversely, IRAS05568-6753 is a C-rich star but was wrongly classified into Oddball because its spectrum not only has C-rich-star features but also has emission in the long IRS wavelength region. In other words, these two categories are not independent from each other. Similarly, six O-rich stars in the ground truth were wrongly classified as C-rich stars by our method. One reason is that the C-rich stars and O-rich stars have a common pattern, in that both have emission in the IRS short wavelength region, although the peaks are not at the same position. The other reason is that the 18-μm-emission peak of these O-rich stars -- the spectral region where there is the most difference between O-rich stars and C-rich stars -- is so weak that the classifier considers it as a small fluctuation, as shown in Fig.2. Consequently, most LMC spectra classification errors were produced due to the interdependency among the C-rich, O-rich and Oddball categories.

Linkage methods are important to the hierarchical clustering. It was proved in [7] and [8] that complete linkage clustering is less sensitive to sources of noise than single linkage clustering, which is also called the nearest neighbor clustering. It was also demonstrated in [9] that average linkage clustering outperforms the other grouping methods. We compared the classification accuracy of two linkage methods (Table 2) and confirmed that average linkage is the best grouping method. Consequently, it should be the first choice when we consider the linkage options to group the infrared spectra.

Table 2 shows that the various scaling methods have a great effect on overall classification accuracy; a good scaling method can improve the classification performance while a bad one always degrades the classification. Fig.3 illustrates that the different scaling methods can change the distance between the objects so as to affect the similarities between the groups and eventually influence the classification results. For instance, Fig.3 (a) and (b) show that the Oddball class is mixed with the other classes using scaling method 2 but it is well separated with the other classes employing scaling method 1. However, the selection of scaling method has to be determined through experiments. In the case of our LMC Spitzer IRS spectral hierarchical clustering, scaling method 2 is superior to the others for either linkage method because it produced the best overall classification accuracy. On the other hand, Table 2 also shows that scaling method 3 led to the best K-means classification result (Fig.3 (d)).

The greatest advantage of PCA is that only a fewer PCs can faithfully represent the data with many dimensions. In our experiment, we reduced the data dimensions from 375 to 3 and only the first three PCs were employed in the procedure. But the space formed by the first three PCs may not be the optimum subspace for classification algorithms to classify the clusters. For instance, the use of the first three PCs degraded the overall classification accuracy of the hierarchical clustering methods as shown in Table 2. The optimum subspace can be found via Multiple Discriminant Analysis [10].

Table 2 demonstrates that the best overall classification accuracy of the hierarchical clustering algorithm, 86.8%, has been achieved under the condition of applying average linkage on the spectra scaled by the maximum amplitude. Meanwhile, the best

overall classification accuracy using the K-means algorithm is 81.1%, lower than the hierarchical algorithm. We caution that this comparison only applies to the LMC Spitzer IRS data set studied here. Compared to the hierarchical algorithm, the drawback of the K-means algorithm is that it requires the start points, as well as the number of classes, as its initial parameters. In our experiment, the selection of start points had a large effect on the final clustering result, because of the small size of our test data set. As the number of samples becomes increasingly larger, sufficient iterations of the K-means algorithm would eliminate the artifacts caused by start points.

**Table 2. The performance comparison of unsupervised clustering algorithms.**

| Unsupervised Clustering Methods | | | | Overall clustering accuracy |
|---|---|---|---|---|
| PCA | Hierarchic clustering | Average Linkage | Scaled data 1 | 73.6% |
| | | | Scaled data 2 | 84.9% |
| | | | Scaled data 3 | 67.9% |
| | | Complete Linkage | Scaled data 1 | 64.2% |
| | | | Scaled data 2 | 79.2% |
| | | | Scaled data 3 | 75.5% |
| | K-means | | Scaled data 1 | 77.4% |
| | | | Scaled data 2 | 71.7% |
| | | | Scaled data 3 | 81.1% |
| Without PCA | Hierarchic clustering | Average Linkage | Scaled data 1 | 81.1% |
| | | | Scaled data 2 | 86.8% |
| | | | Scaled data 3 | 75.5% |
| | | Complete Linkage | Scaled data 1 | 71.7% |
| | | | Scaled data 2 | 83.0% |
| | | | Scaled data 3 | 83.0% |
| | K-means | | Scaled data 1 | 77.4% |
| | | | Scaled data 2 | 71.7% |
| | | | Scaled data 3 | 81.1% |

***Note**: Scaled data 1, 2 and 3 represent the spectral data set scaled by scaling method 1, 2, and 3, respectively.*

## 4. Conclusion

Hierarchical clustering and K-means clustering were implemented to classify our test data set of Spitzer IRS spectra of LMC infrared sources. The effects of some factors, such as linkage method, data rescaling and PCA have been investigated. The best overall classification accuracy, 86.8% (or the least classification error, 13.2%) has been achieved by the hierarchical clustering algorithm with the combination of average linkage and the data scaled according to maximum amplitude. PCA was used to reduce the dimensionalities of LMC spectra. The first three PCs, accounting for more than 95% of total variance, were employed for hierarchical clustering and K-means clustering. The cost of the use of the first three PCs is that it decreased the overall classification accuracy of hierarchical clustering algorithms in our experiment.

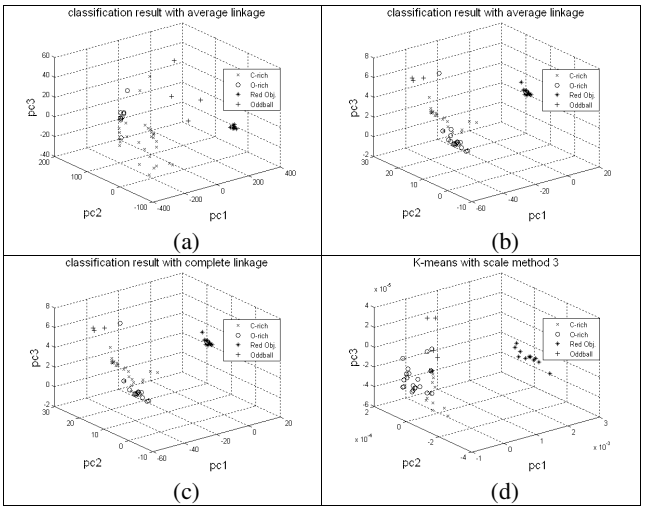


***Figure 3**. Classification results: (a). average linkage + scaling method 1; (b). average linkage + scaling method 2; (c). complete linkage + scaling method 2; (d). K-means + scaling method 3. **Note:** ' x ' refers to C-rich stars; ' o ' refers to O-rich stars; ' * ' refers to Red Objects; ' + ' refers to Oddballs.*

## References


[1]. H. van Winckel, "Post-Agb Stars", Annual Review of Astron. and Astrophys., 41, 391 (2003).

[2]. C. L. Buchanan, J. H. Kastner, et al, "A Spitzer IRS Spectral Atlas of Luminous 8 μm Sources in Large Magellanic Cloud", submitted to Astron. J.

[3]. B. S. Everitt, Cluster Analysis, 3rd ed (Oxford Univ. Press Inc, 1993) pg 55.

[4]. J. E. Jackson, A User's Guide to Principal Components (John Wiley & Sons, Inc. 1991)pg 10.

[5]. R. G. Congalton, "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data", Remote Sensing of Environment, 37, 35 (1991).

[6]. J. H. Kastner, C. L. Buchanan, et al, "Spitzer Spectroscopy of Dusty Disks around B[e] Hypergiants in the Large Magellanic Cloud", Astrophys. J. (letters), in press.

[7]. F. B. Baker, "Stability of Two Hierarchical Grouping Techniques--Case 1. Sensitivity to Data Errors", J. Amer. Statistic Assoc., 69, 440 (1974).

[8]. L. J. Hubert, "Approximate Evaluation Techniques for the Single Link and Complete Link Hierarchical Clustering Procedures", J. Amer. Statistic. Assoc., 69, 698 (1974).

[9]. K. M. Cunningham, J. C. Ogilive, "Evaluation of Hierarchical Grouping Techniques: A Preliminary Study", Comp. J., 15,209 (1972).

[10]. R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd ed (John Wiley, Inc. 2000)pg 121.


## Author Biography

*Bo Mu received his B.S. degree in Mechanical Engineering and M.S. degree in Computer Graphics from Beijing Institute of Technology in 1994 and 1998, respectively. He is currently pursuing his PhD degree in Imaging Science at Rochester Institute of Technology. Since 2003, he has been working on developing an unsupervised pattern classification algorithm on X-ray CCD spectra with Professor Joel Kastner. His research interests include digital image processing, digital video processing, pattern classification and X-ray astronomy.*