CLASSIFICATION OF ASTRONOMICAL INFRARED SOURCES USING SPITZER SPACE TELESCOPE DATA

MICHAEL DENNING

SENIOR RESEARCH: FINAL REPORT

ADVISORS: DR. CARL SALVAGGIO, DR. JOEL KASTNER

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE

ROCHESTER INSTITUTE OF TECHNOLOGY

MARCH 2007

ABSTRACT

NASA's Spitzer Space Telescope is generating thousands of spectra of mass-losing stars in the Milky Way and nearby galaxies. These mass-losing stars allow researchers to determine where and how the fundamental elements of the universe first came to be. By extracting spectral source fluxes of objects imaged with the Spitzer Space Telescope (SST), it is possible to classify mass-losing objects in galaxies using the various emissive properties and characteristics inherent to these entities. First, however, one must develop and evaluate the effectiveness of classification strategies and methods for these objects.

This research project is aimed at classifying mass-losing stars located in the nearby Large Magellanic Cloud (LMC) galaxy into subgroups according to their composition. To classify a group of spectra of objects in the LMC, this technique first employs a hierarchical clustering routine to combine similar spectra within the normalized dataset, followed by a K-means clustering algorithm to refine the groups and thereby provide a more accurate classification.

A classification technique for X-ray sources observed with the Chandra Space Telescope was originally developed by Hojnacki *et al.* (2007). We found that implementing this routine on SST infrared spectra allowed for the classification of a known group of stars into seven subdivisions with an accuracy of approximately 80%. Using the spectral signatures from the classification on the known initial dataset (Buchanan *et al.* 2006), this technique was applied to a much larger set of stars of unknown IRS type (Sloan, unpublished). While we found that the technique accurately groups similar spectra, the exact classification accuracy for this dataset remains undetermined as astronomers continue to study and understand the nature of these objects.

As a result of this research, we have shown that it is indeed possible to classify and group Spitzer IRS spectra of astronomical objects with minimal initial information. Using this technique, astronomers and scientists alike now have a powerful tool at their disposal that may be used to classify large datasets of astronomical spectra.

TABLE OF CONTENTS

BACKGROUND

Stars and Spectroscopy with the Spitzer Space Telescope

Stars undergo a well-defined set of evolutionary steps over their lifetime (University of Utah 2006). The majority of a star's lifecycle is spent in the period known as the main sequence (MS). During this period, stars undergo hydrogen fusion and ultimately emit this fusion-generated energy into space. Traditionally, this energy has been measured through optical spectroscopy, enabling astronomers to study the composition and development of the star during the main sequence phase. When hydrogen in a star runs out, the star enters the post-MS phase as fusion halts and the temperature drops. The core then contracts ultimately causing the temperature to rise again. The high density in the core causes nuclear fusion to begin again, this time with helium being converted to carbon. The star becomes less and less stable as a larger amount of energy is emitted. When helium runs out and the star is sufficiently massive, the core contracts again and reinitiates fusion with carbon. Over time, the products of fusion can be mixed to the star's surface. As its post-MS evolution continues, the star becomes surrounded by thick dust that absorbs visible-wavelength emission and radiates the energy in the infrared region. Therefore, astronomers study this post-MS period by imaging stars in the infrared region.

The Spitzer Space Telescope, one of NASA's "great observatories," was launched in 2003 as an infrared space observatory. Three main instruments are on board, including the Infrared Array Camera (IRAC), Infrared Spectrograph (IRS), and the Multiband Imaging Photometer for Spitzer (MIPS). This study involves data acquired with the IRS.

The IRS has four main modules, including low spectral resolution ($E/\Delta E \approx 200$) instruments sensitive from 5.3μm - 14μm and 14μm - 40μm, and two high spectral resolution

($E/\Delta E \approx 600$-1000) instruments that operate from 10μm – 19.5μm and 19μm - 37μm. Since many bright infrared objects, many of which are mass-losing stars, exist in the nearby Large Magellanic Cloud (LMC) galaxy, their spectra were captured using the Spitzer Space Telescope IRS (Buchanan *et al.* 2006; Sloan, unpublished). Each object has unique characteristics that may be analyzed by the energy that is emitted at various infrared wavelengths. For each object, these energies produce a spectrum that is characteristic of specific gases and solid materials (dust grains). Astronomers may study these emissive properties to learn more about the star's development, lifecycle, and eventual death.

Due to the large number IRS spectra obtained with the SST, it is not always practical or efficient to examine each and every spectrum. Classification methodologies would greatly benefit the astronomical community by grouping similar spectra into subgroups. Under the assumption that each object that is a member of the same subgroup will have similar spectral properties, its emissive properties only need to be classified once. The resulting database of classified spectra would provide astronomers with valuable information, allowing them to examine how fast these stars are losing mass, and at what total rate elements such as carbon and oxygen are being returned to space. This knowledge would allow for a better understanding of how carbon-based life forms, such as those here on Earth, came into existence and whether we are likely to find life-enabling substances elsewhere in the universe.

Clustering Techniques

Unsupervised classification methods are useful when there is not sufficient training data or the number of spectral classes is unknown. Clustering may be used as a method to achieve unsupervised classification as it allows for the classification of objects into groups, where each group contains objects that are similar in *n*-dimensional space (Richards 1999). Clustering

procedures can be used to segment a spectrum (or any dataset of *n*-dimensional vectors) into unknown classes that may later be identified. The procedure involves grouping a set of data points in a highly dimensional space so that the points that are members of a particular cluster are also spectrally similar. To classify a set of astronomical spectra, each spectrum may be considered as a vector of *n* points. Clustering procedures may then be initiated on the spectral database so that the database will be subdivided into groups containing similar spectra. The resulting clusters may then be classified based upon inherent spectral properties of known star composition types.

Hierarchical Clustering

The two primary types of clustering are hierarchical and nonhierarchical (partitional). Hierarchical methods determine successive clusters based upon an initial or previous cluster, and may operate in an agglomerative or divisive nature (Duda 2001). Agglomerative clustering begins with each point as its own cluster, and then combines points based on their similarity into larger clusters. Divisive clustering begins with all data points in one cluster and successively breaks them apart into smaller clusters based on some distance metric or measure of similarity. The results of both agglomerative and divisive clustering can be displayed in a two-dimensional figure known as a dendrogram, as shown below in Figure 1.

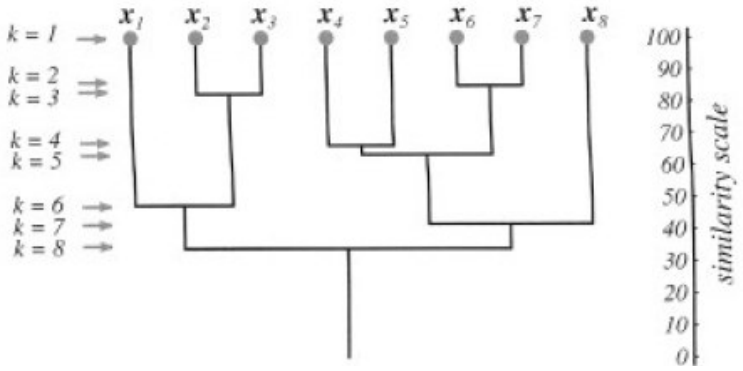

Figure 1. Dendrogram representing similarity metric among clusters of a hierarchical clustering routine (Duda 2001)

Figure 1 represents the results of hierarchical clustering. The vertical axis indicates a generalized measure of similarity among clusters. At level 1, all eight objects reside in their own cluster, as each object is highly similar to itself. Points $\mathbf{x}_6$ and $\mathbf{x}_7$ are the next two most similar clusters, and so they are merged at level 2. The process repeats until all objects reside in one cluster (Duda 2001).

When performing agglomerative clustering, the "linkage method" determines how clusters are merged. These three linkage types are portrayed in Figure 2.

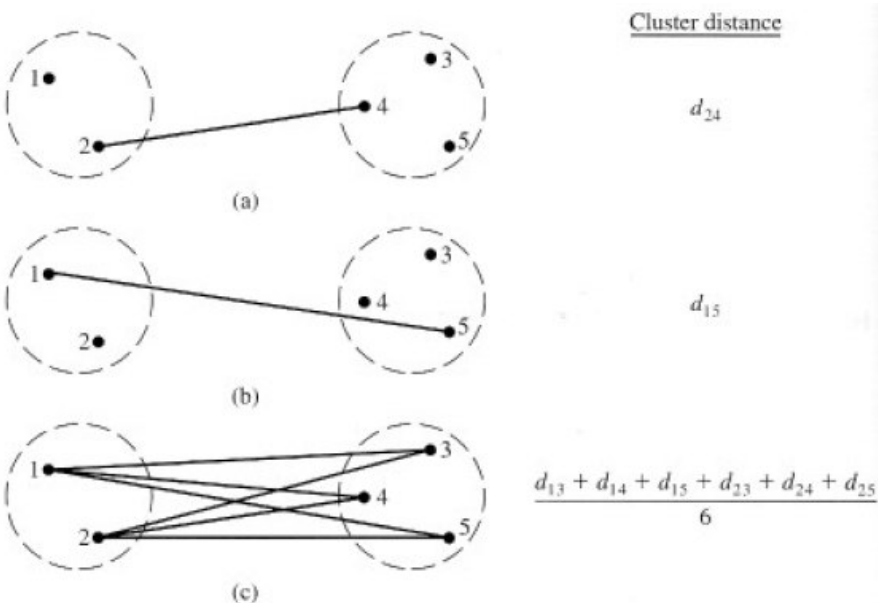

Figure 2. Linkage methods for agglomerative clustering: (a) single linkage, (b) complete linkage, and (c) average linkage (Johnson 2002)

Single linkage uses the minimum distance, complete linkage uses the maximum distance, and average linkage uses the average distance between all pairs of points.

According to Johnson (2002), the steps to agglomerative hierarchical clustering for grouping N objects are as follows:

1. Start with N clusters, each containing a single entity, and an N x N symmetric matrix of cluster distances (or similarities, as defined by the linkage method described in Figure 2).

2. Search the distance matrix for the nearest (or most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{UV}.
3. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters U and V and (b) adding a row and column giving the distances between cluster (UV) and the remaining clusters.
4. Repeat steps 2 and 3 a total of N-1 times. All objects will be in a single cluster after the algorithm terminates. Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

While many similarity metrics are acceptable for step one, the Euclidean distance is the most common distance measurement used to compare distance between each pair of spectra (or cluster) in the dataset. This distance is defined for multi-dimensional data as

$$d(\mathbf{u}_i, \mathbf{v}_k) = \sqrt{(\mathbf{u}_i - \mathbf{v}_k)^t (\mathbf{u}_i - \mathbf{v}_k)} \qquad (1)$$

where $\mathbf{u}_i$ and $\mathbf{v}_k$ are cluster vectors containing elements $i = 0, 1, \ldots M$ and $k = 0, 1, \ldots M$, where M is the number of elements in each vector. To complete the N x N similarity matrix, the distance between all possible pairs of clusters ($\mathbf{u}_i$, $\mathbf{v}_i$) is calculated according to Equation 1. The two clusters that exhibit the least distance are then merged into one cluster based upon the linkage method. Step four indicates that the procedure repeats until all objects are in a single cluster. However, the procedure may be terminated at any point defined by the user, whether it be a specific number of clusters or level of similarity among them. Once agglomerative clustering is achieved, the result is a group of N clusters, each with member(s) that are similar to each other in spectral space.

K-means Clustering

The K-means algorithm, unlike hierarchical clustering methods, requires the user to set the number of clusters and define an initial "seed" for each cluster. The class of a spectral vector is initially determined by calculating which cluster seed is closest to the vector by some distance metric, often the Euclidean distance. A matrix of similarities is not required, thus nonhierarchical techniques such as K-means may be applied to larger datasets than hierarchical methods. The K-means algorithm starts with *K* initial vectors $\hat{\boldsymbol{\mu}}_i$, representing each cluster mean

$$\hat{\boldsymbol{\mu}}_i \ \ where \ \ i = 1,2,...K \qquad (2)$$

The seeds are chosen either randomly, as points from the original dataset, or by known characteristics of the expected clusters. Each object **x** is then classified as a member of the nearest cluster based upon the least Euclidean distance from the cluster mean to that object, defined as

$$\mathrm{d}(\mathbf{x}, \hat{\boldsymbol{\mu}}_i) = \sqrt{(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^t (\mathbf{x} - \hat{\boldsymbol{\mu}}_i)} \qquad (3)$$

Once each object has been classified, the cluster seed is replaced by the mean of the members of that particular cluster and represented as

$$\boldsymbol{\mu}_i \ \ where \ i = 1,2,...K \qquad (4)$$

Once new cluster means are computed, the procedure is repeated until a terminator threshold is met. The complete steps involved in K-means clustering for grouping N objects are as follows:

1. Define the number of clusters, K.
2. Set an initial cluster seed for each cluster K, defined randomly from the dataset or by known inherent cluster properties.
3. Examine all objects, assigning each object to the cluster whose seed or centroid (mean) is nearest, based upon some distance metric.

4. Once all objects have been assigned, recalculate cluster means.
5. Repeat steps 3 and 4 until a.) no additional reassignments occur, or b.) a pre-determined reassignment threshold is met.

It is often not practical to repeat clustering until the change in cluster means is zero. Instead, the difference of these "migrating means" can be tracked for each iteration, and the K-means procedure can be terminated when this difference falls below a threshold, often defined as a percent error between zero and five percent.

The difference in migrating means and the effect of each clustering iteration is often examined by analyzing the progressive reduction in the sum of squared error (SSE) between successive iterations. The SSE between cluster vectors before and after an iteration is defined as

$$SSE = \sum_{C_i} \sum_{x \in C_i} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i)^t (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i) \qquad (5)$$

where the outer summation is the sum across all clusters, and the inner summation is across all objects in the i^{th} cluster. The cluster mean before the iteration is defined as $\hat{\boldsymbol{\mu}}_i$, and the new computed cluster mean vector after the iteration is represented by $\boldsymbol{\mu}_i$. When the percent change in SSE between iterations falls below a user-defined threshold, the procedure will be terminated. The result is a cluster map defining class membership for each object in the dataset.

Data

Infrared spectra for 51 luminous stars in the LMC were obtained from a recent study by Buchanan *et al.* (2006). These data have a spectral range of approximately 5.0 to 35.0 microns in wavelength, and each spectrum contains 300-400 samples which provide the corresponding flux (in units of Janskys, where 1.0 Jy = 10^{-23} erg/(cm^2·sec·Hz)) for each wavelength.

Sample IRS spectra are provided in Figure 3 for seven objects. Each spectrum is characteristic of specific elemental (chemical) compositions and/or stellar evolutionary states,

and is representative of one of seven unique classes or IRS types as defined by Buchanan *et al.*. By comparing the database of objects with reference spectra such as these, classifications of other IRS spectra may be made.

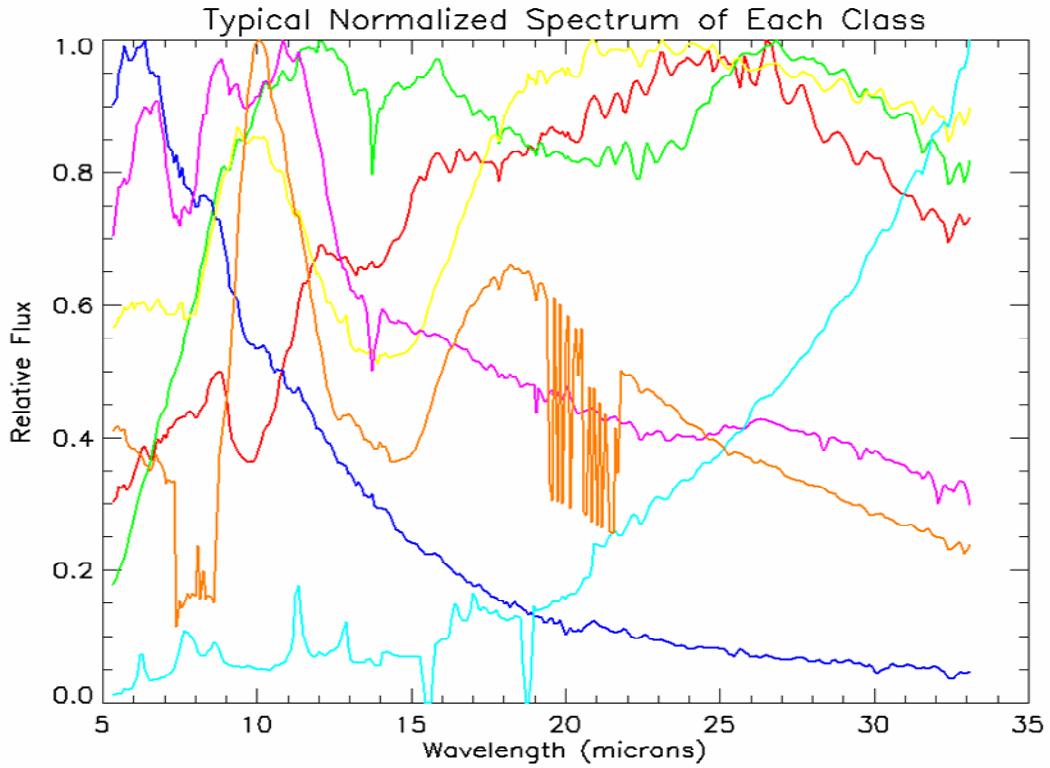

Figure 3. Sample normalized spectra of each IRS type from the Buchanan database, including O AGB (green), RSG (orange), MW O AGB (blue), OH/IR SG (red), C AGB (purple), HII (cyan), Peculiar (yellow).

Figure 3 depicts sample spectra for each of seven IRS types that correspond to the chemical composition surrounding the star, as defined by Buchanan *et al.*: O AGB (oxygen-rich ejecta), RSG (red supergiant, oxygen rich), MW O AGB (Milky Way star), OH/IR SG (oxygen rich), C AGB (carbon-rich ejecta), HII ("red" spectra displaying polycyclic aromatic hydrocarbon [PAH]), and "peculiar" (oxygen rich, used to classify those spectra that are not similar to any of the other six IRS types). Aside from MW O AGB, all of these stars reside in the LMC.

These 51 objects provide excellent overall coverage of luminous infrared objects often imaged by the Spitzer Space Telescope. While Buchanan *et al.* have classified each of the 51

stars into one of the above IRS types, for our purposes classification will proceed blindly. Once the clustering routines pair each spectrum with an IRS type, we will then refer back to Buchanan *et al.* as a reference for determining classification accuracy.

EXPERIMENTAL METHOD

The experimental method consists of five main parts. First, we will extract source flux information for each object. Once the spectra have been extracted, we will peak normalize and interpolate each spectrum so that all spectra have relative flux densities over identical wavelengths. Next, classification will begin by applying an agglomerative hierarchical clustering algorithm. The procedure will repeatedly iterate until only seven clusters exist. Next, we will calculate cluster means for each cluster as obtained from the hierarchical method, and complete K-means clustering using the cluster means as seeds. K-means will repeat until no reassignments occur. With a class membership table, we will analyze and determine the classification accuracy of the Buchanan dataset. Lastly, the above procedure will be repeated on a second group of objects captured by the SST, the Sloan database. By using the results of hierarchical clustering on the Buchanan database as initial cluster seeds for K-means processing on the Sloan database, we will examine the classification accuracy of these objects of unknown IRS type in the Sloan database.

Preparing the Data

Spectral data were stored in the FITS format, a data file type commonly used in the astronomical community. Using IDL, source flux information and corresponding wavelengths were extracted from each of the 51 FITS files. Because the range and number of wavelength samples were different across each spectrum, the spectra were interpolated so that each contained 325 points from 5.34μm to 33.09μm, with uniform spectral spacing of 0.0854μm. We ignored flux information outside our chosen wavelength range, which was selected so that flux information existed for all points within the selected range.

Once the data were interpolated, we had 51 spectral vectors, each with 325 samples. Seven of these spectral vectors representing each of the IRS types are shown in Figure 4.

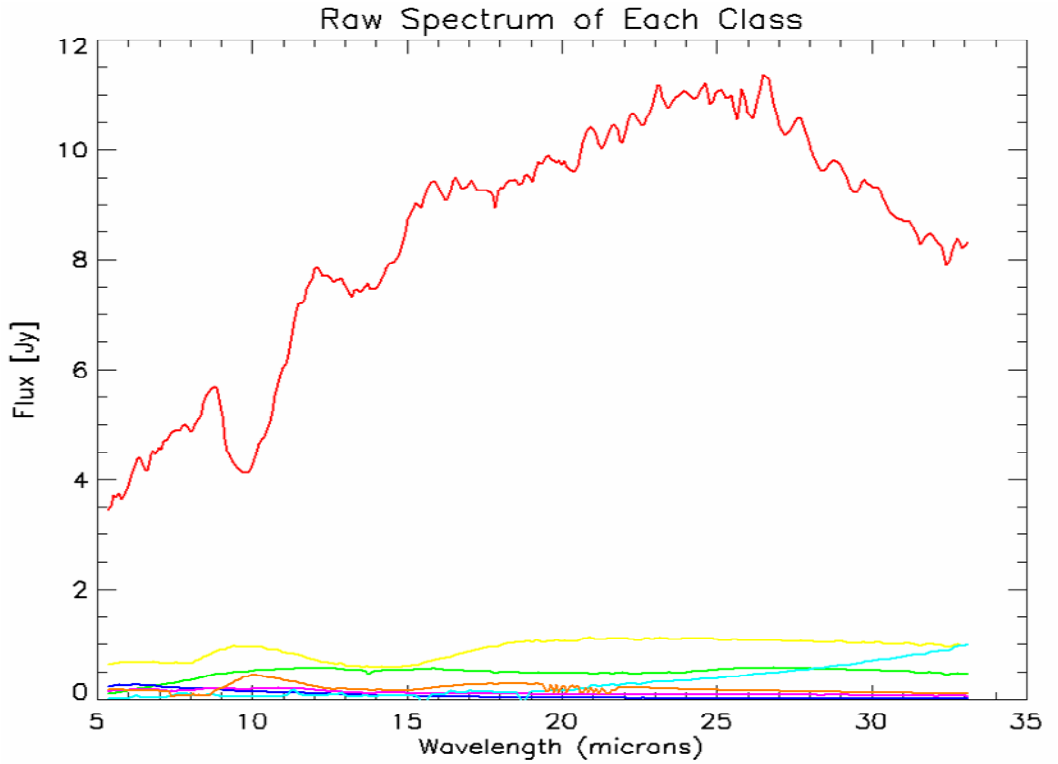

Figure 4. Interpolated, unnormalized spectra of each IRS type from the Buchanan database, including O AGB (green), RSG (orange), MW O AGB (blue), OH/IR SG (red), C AGB (purple), HII (cyan), Peculiar (yellow).

Figure 4 indicates the difference in flux density for typical spectra of various IRS types. C AGB stars, for example, have spectra below a flux density of 1 Jy, and are dominated by noise at points. OH/IR SG type stars can have a flux density as high as about 10 Jy. In an effort to positively influence clustering results, all spectra were scaled by the maximum amplitude of the spectra as shown in Figure 3. Lastly, any samples with a flux below zero were set equal to zero so that all spectra exhibit relative flux densities between 0.0 and 1.0.

Agglomerative Hierarchical Clustering

With 51 interpolated and scaled spectra representing each of 51 individual clusters, we began to apply the hierarchical clustering technique by calculating the Euclidean distance, as

defined by Equation 1, for each pair of spectra in the dataset. Once the two spectra separated by the least distance were determined, the two were merged into single cluster. A study by Mu *et al.* examined the three linkage methods as described Figure 2. Since Mu *et al.* demonstrated that average linkage provided the best clustering results, we chose to merge the two most similar spectra using this metric.

The clustering routine was repeated, reducing the total number of clusters by one for each iteration. Each spectrum from the Buchanan database was classified into one of seven groups; as a result, we chose to terminate the hierarchical clustering algorithm at the level containing seven clusters. Figure 5 supports our choice of seven clusters.

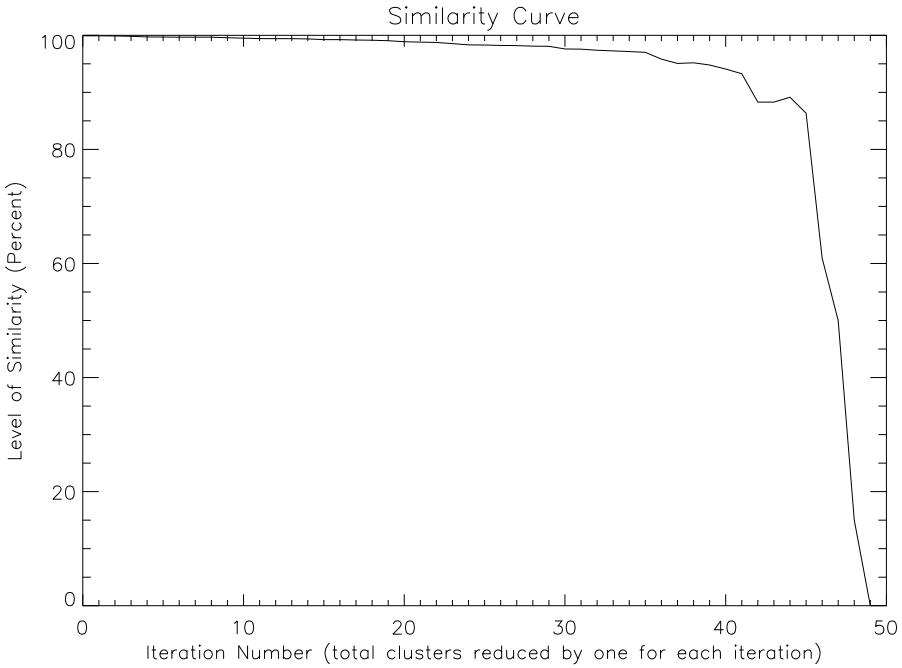

Figure 5. The similarity curve shows how the overall level of similarity among clusters drops as the clustering routine iterates and the number of clusters is successively reduced by one. There is a sharp drop in level of similarity at iteration 45 where the number of clusters is 7.

The similarity between merged clusters, provided as a scaled sum of square error between the two clusters, drops off significantly at iteration 45. Termination of the clustering algorithm at iteration number 45 results in seven remaining clusters. Any successive iterations would

combine clusters that do not share similar spectra. If the procedure was terminated prior to iteration 45, clusters with spectrally similar data would be left unmerged. For these reasons, we chose seven as the ideal number of clusters.

K-means Clustering

As a result of hierarchical clustering, all 51 spectra were classified into one of seven subgroups. In an effort to increase classification accuracy, we chose to implement K-means clustering using initial cluster means as defined by the results of hierarchical clustering. The K-means algorithm then reassigned spectra based on the closest cluster mean.

The implementation of K-means followed the guidelines discussed on page 10. As indicated, the initial cluster means representing each of the seven clusters were obtained by averaging all spectral vectors in each of the clusters that resulted from hierarchical clustering. The 51 spectra were then examined on an individual basis; the Euclidean distance between the spectrum and each of the seven initial cluster means was calculated, and that spectrum was classified as part of the cluster which resulted in the least Euclidean distance. After all 51 spectra were examined, new cluster means were calculated, and the process iterated until cluster membership no longer changed from iteration to iteration. The result is a final list of cluster membership for each of the 51 spectra.

Classification Accuracy

Once we obtained a list of the seven classes and the objects belonging to each, we examined the spectra and labeled the IRS type corresponding to each cluster using the guidelines as discussed in Buchanan *et al*. A confusion matrix of the results was created, and the overall classification accuracy was determined. Various spectra and their class membership were analyzed to examine how well the classification routine worked.

Sloan Database

Lastly, we applied K-means clustering (using initial cluster means from hierarchical clustering of the Buchanan dataset) to another database of infrared spectra in the LMC. This dataset, provided by G. Sloan (unpublished), contained 184 objects. These spectra were peak normalized and interpolated so that they contained the same number of samples and points as during our classification implementation of the Buchanan spectral database.

Instead of reapplying hierarchical clustering to the Sloan database, we chose to simply use the results from hierarchical clustering of the Buchanan spectra. The seven IRS types have distinctive spectra, and it is not necessary to perform hierarchical clustering on the Sloan database to obtain a new set of initial cluster means. Instead, we used the same initial cluster means and classified each of the 184 objects from Sloan through use of K-means clustering. The majority of these 184 objects, however, have not yet been classified as members of a particular IRS type by the astronomical community. Exact classification accuracy, therefore, cannot be obtained at the present time. Instead, we compared the spectral shapes that corresponded to each class of the Buchanan and Sloan databases and drew preliminary conclusions as to how well our classification technique performs when blindly implemented on a database of infrared spectra of objects with unknown classes.

RESULTS & DISCUSSION

The resulting dendrogram from hierarchical clustering on the Buchanan dataset is shown in Figure 5.

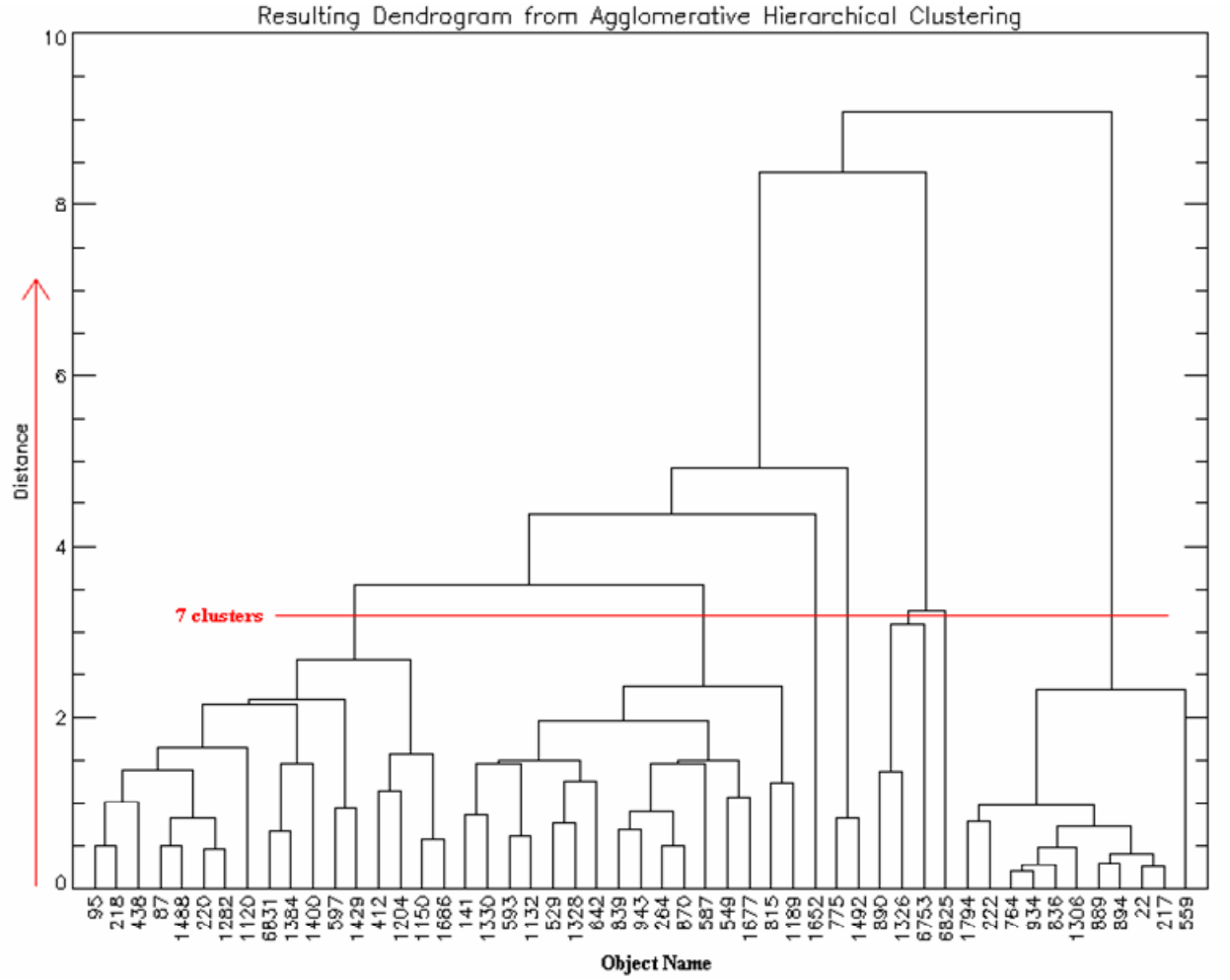

Figure 5. The dendrogram shows the division of clusters as a result of hierarchical clustering on the Buchanan database. The horizontal red line depicts the level at which only seven clusters exist; the vertical red line indicates the successive reduction in number of clusters as the "distance" among them increases.

Moving up from the bottom, we can see that the most similar clusters are continually merged. The horizontal red line indicates the point at which the algorithm was terminated and seven clusters were reached. HII objects have a distinctive spectrum unlike the other types, and as expected, the eleven objects on the right edge of the dendrogram indicate that these are among

the first spectra to be combined. We can also see that two "clusters," OH/IR SG and O AGB, consist of only one object. MSXLMC890, MSXLMC1326, and IRAS05568-6753 were classified in the same group, yet IRAS05568-6753 is not a "peculiar" object. In an effort to increase classification accuracy and "push over" objects exhibiting subtle differences such as these, K-means clustering was applied. A total of eight iterations were needed before no reassignments occurred. After eight iterations were completed, classification was complete and the resulting class membership table for all objects is provided below. As a result of implementing K-means clustering, four objects changed cluster membership originally assigned by the hierarchical method.

Table 1. Class membership table for the Buchanan spectral database as result of implementation of the classification scheme, including hierarchical and K-means clustering. Green depicts a correct classification, while red indicates an incorrect classification.

Object Name	Actual IRS Type	Classified IRS Type	Correct?	Object Name	Actual IRS Type	Classified IRS Type	Correct?
MSXLMC87	C AGB	C AGB		MSXLMC412	MW O AGB	C AGB	
MSXLMC95	C AGB	C AGB		MSXLMC1150	MW O AGB	MW O AGB	
MSXLMC218	C AGB	C AGB		MSXLMC1677	MW O AGB	RSG	
MSXLMC220	C AGB	C AGB		MSXLMC1686	MW O AGB	MW O AGB	
MSXLMC438	C AGB	C AGB		MSXLMC642	O AGB	RSG	
MSXLMC775	C AGB	MW O AGB		IRAS04553-6825	OH/IR SG	OH/IR SG	
MSXLMC1120	C AGB	C AGB		MSXLMC890	Peculiar	Peculiar	
MSXLMC1282	C AGB	C AGB		MSXLMC1326	Peculiar	Peculiar	
MSXLMC1384	C AGB	C AGB		MSXLMC141	RSG	RSG	
MSXLMC1400	C AGB	C AGB		MSXLMC264	RSG	RSG	
MSXLMC1488	C AGB	C AGB		MSXLMC529	RSG	RSG	
MSXLMC1492	C AGB	MW O AGB		MSXLMC549	RSG	C AGB	
IRAS05568-6753	C AGB	O AGB		MSXLMC587	RSG	RSG	
MSXLMC1652	C AGB	C AGB		MSXLMC593	RSG	RSG	
IRAS04374-6831	C AGB	C AGB		MSXLMC597	RSG	C AGB	
MSXLMC22	HII	HII		MSXLMC815	RSG	RSG	
MSXLMC217	HII	HII		MSXLMC839	RSG	RSG	
MSXLMC222	HII	HII		MSXLMC870	RSG	RSG	
MSXLMC559	HII	HII		MSXLMC943	RSG	RSG	
MSXLMC764	HII	HII		MSXLMC1132	RSG	RSG	
MSXLMC836	HII	HII		MSXLMC1189	RSG	RSG	
MSXLMC889	HII	HII		MSXLMC1204	RSG	C AGB	
MSXLMC894	HII	HII		MSXLMC1328	RSG	RSG	
MSXLMC934	HII	HII		MSXLMC1330	RSG	RSG	
MSXLMC1306	HII	HII		MSXLMC1429	RSG	C AGB	
MSXLMC1794	HII	HII					

Table 1 provides the final class membership list after the clustering classification technique was implemented. The single OH/IR SG object, both peculiar objects and all eleven HII objects were classified correctly. The single O AGB object, however, was incorrectly classified as RSG. Twelve of fifteen C AGB objects were classified correctly; the remaining three were classified as either O AGB or MW O AGB. Two of four MW O AGB objects were classified correctly, and thirteen of seventeen RSG objects were correct. Figure 6 illustrates the spectral members of the resulting groups, and Table 2 on the next page represents a confusion matrix of the data and overall classification accuracy statistics.

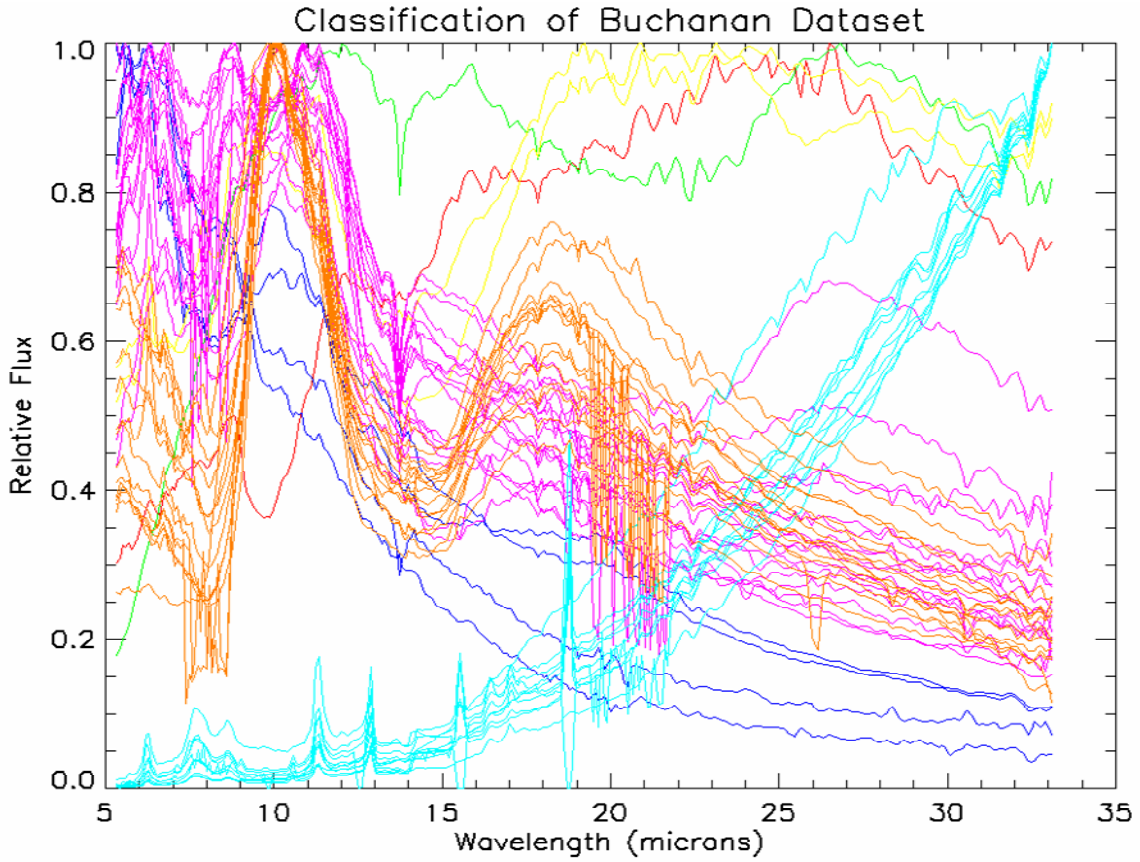

Figure 6. Classification results representing the spectral subdivision of the Buchanan database; IRS types include O AGB (green), RSG (orange), MW O AGB (blue), OH/IR SG (red), C AGB (purple), HII (cyan), Peculiar (yellow).

Table 2. Confusion matrix and overall classification statistics

		Classified							
		C AGB	**HII**	**MW O AGB**	**O AGB**	**OH/IR SG**	**Peculiar**	**RSG**	**Total**
Actual	**C AGB**	12	0	2	1	0	0	0	**15**
	HII	0	11	0	0	0	0	0	**11**
	MW O AGB	1	0	2	0	0	0	1	**4**
	O AGB	0	0	0	0	0	0	1	**1**
	OH/IR SG	0	0	0	0	1	0	0	**1**
	Peculiar	0	0	0	0	0	2	0	**2**
	RSG	4	0	0	0	0	0	13	**17**
	Total	**17**	**11**	**4**	**1**	**1**	**2**	**15**	**51**
Objects classified correctly					**41**				
Objects classified incorrectly					**10**				
Overall classification accuracy					**80 ± 3 %**				
Overall classification error					**20 ± 1 %**				

Our classification scheme accurately classified 80% of the spectral database into the correct IRS type groups. Forty-one objects were classified correctly while ten objects were grouped into an incorrect IRS group. Examination of the spectra, as shown in Figure 6, indicated the similarity among spectra grouped into the same IRS types. The O AGB, MW O AGB, and peculiar spectra all have a relative flux above 0.8 from 19μm to 33 μm. The O AGB spectrum, however, peaks at 12μm after a sharp increase in flux at its minimum near 5μm. At 12μm, the peculiar spectra drop to nearly half the relative flux when compared to the O AGB spectra; for this reason, clustering placed these objects into two different groups. The green spectrum, however, is actually a C AGB type. As illustrated, its spectrum is quite different from the other C AGB objects (purple spectra). The clustering algorithm accurately grouped all HII objects into their own group. Their spectra are shown in Figure 7.

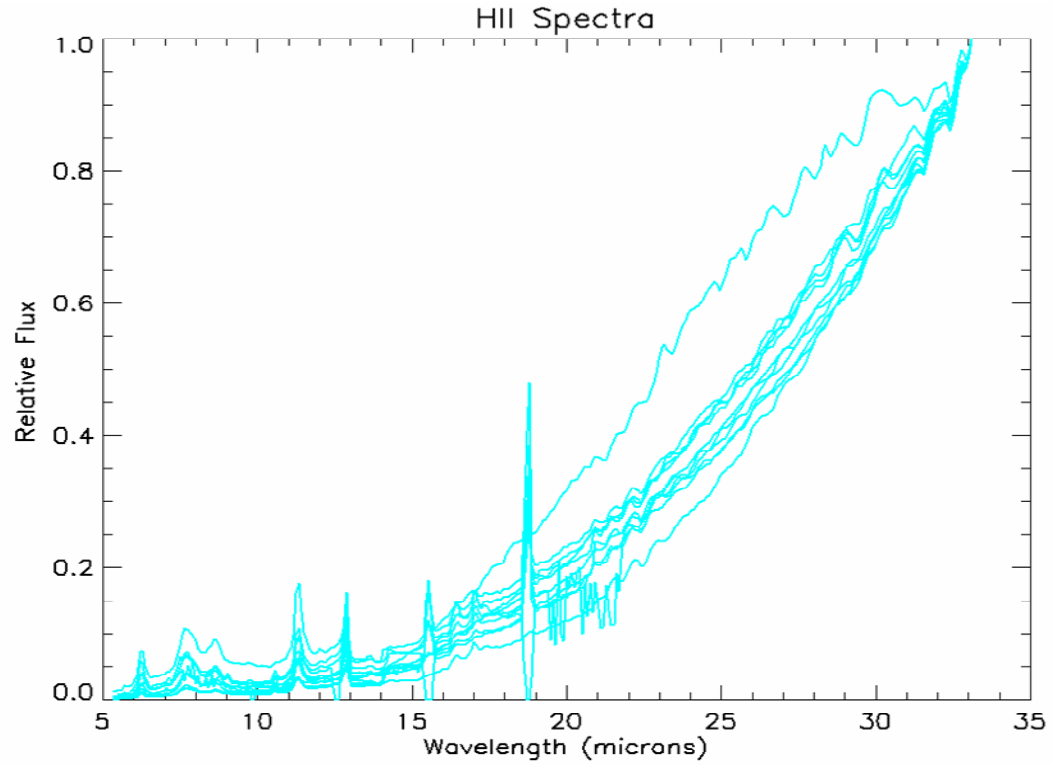

Figure 7. Due to the strong similarity among the HII spectra, the clustering algorithm accurately classified 100% of the HII spectra in the Buchanan database.

As illustrated in Figure 7, all HII objects exhibit very similar spectra. This similarity, combined with their distinction from the other object types, caused the clustering routines to successfully group them together into their own IRS type group. While this is an example where the classification procedure worked well, the technique did not work so well for classifying MW O AGB type objects. The spectra for those objects classified or defined as type MW O AGB are in Figure 8.

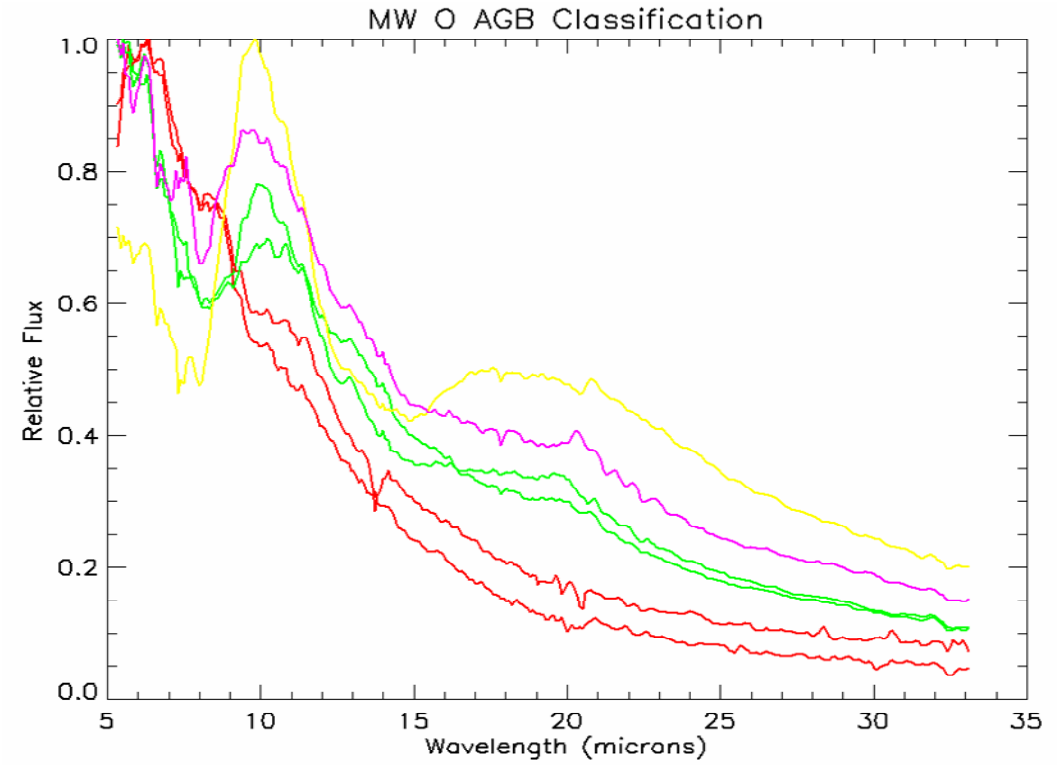

Figure 8. The classification scheme correctly identified only two of the four MW O AGB spectra (green). Two spectra classified as MW O AGB were actually C AGB (red). The two remaining MW O AGB objects were classified as C AGB (purple) and RSG (yellow).

Two of four MW O AGB spectra were classified correctly, as depicted by the green spectra in Figure 8. The other two spectra classified as MW O AGB were actually of type C AGB, depicted by the red spectra. The two actual MW O AGB spectra were classified as members of types C AGB and RSG. The two C AGB spectra, unlike the MW O AGB spectra, do not exhibit a spike at 11μm. Nevertheless, they were classified as MW O AGB because the Euclidean distance between them was less than that of the remaining C AGB type objects. This is perhaps one downside to the clustering technique as it fails to weigh certain known characteristics of specific object types.

Overall, however, the classification technique worked well and classified 80% of the spectral database correctly. To further examine the technique, we applied K-means clustering to

the Sloan spectral database using initial cluster means from hierarchical clustering of the Buchanan database. The resulting spectral classes are shown below.

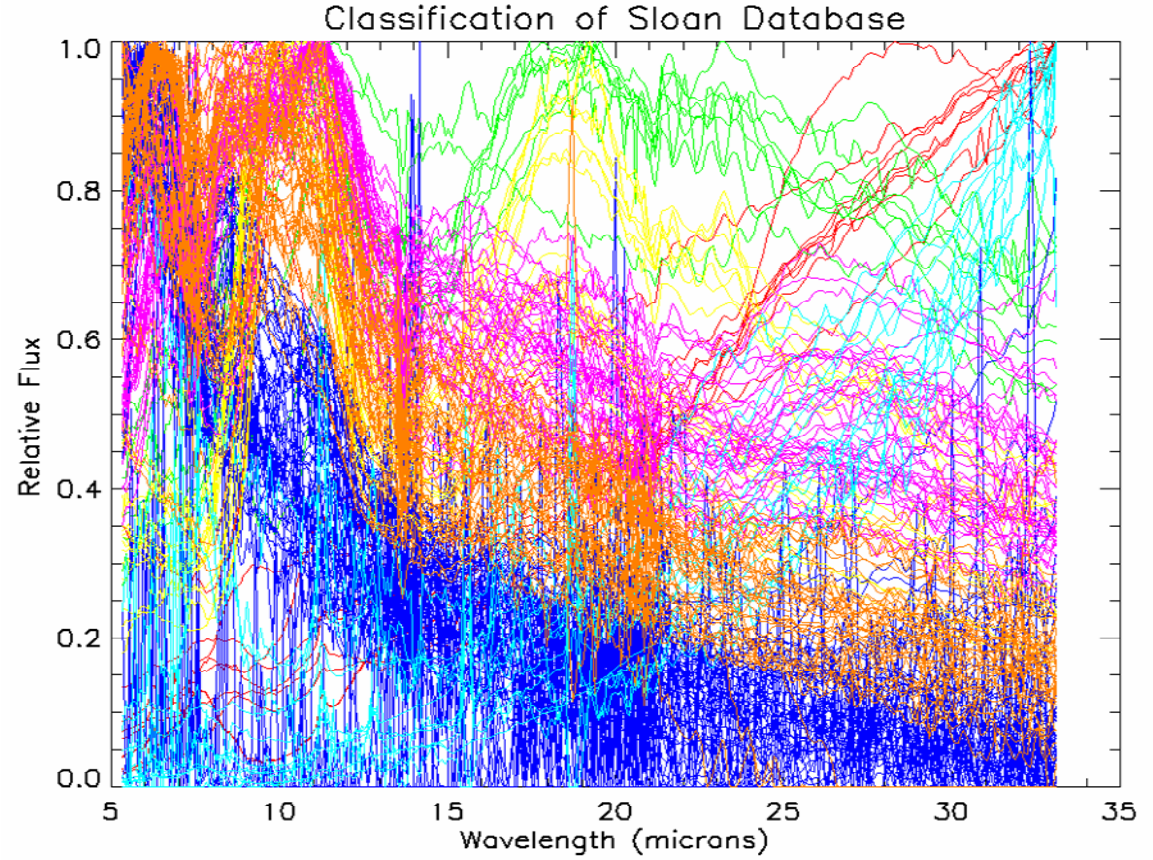

Figure 9. Classification results representing the spectral subdivision of the Sloan database; IRS types include O AGB (green), RSG (orange), MW O AGB (blue), OH/IR SG (red), C AGB (purple), HII (cyan), Peculiar (yellow).

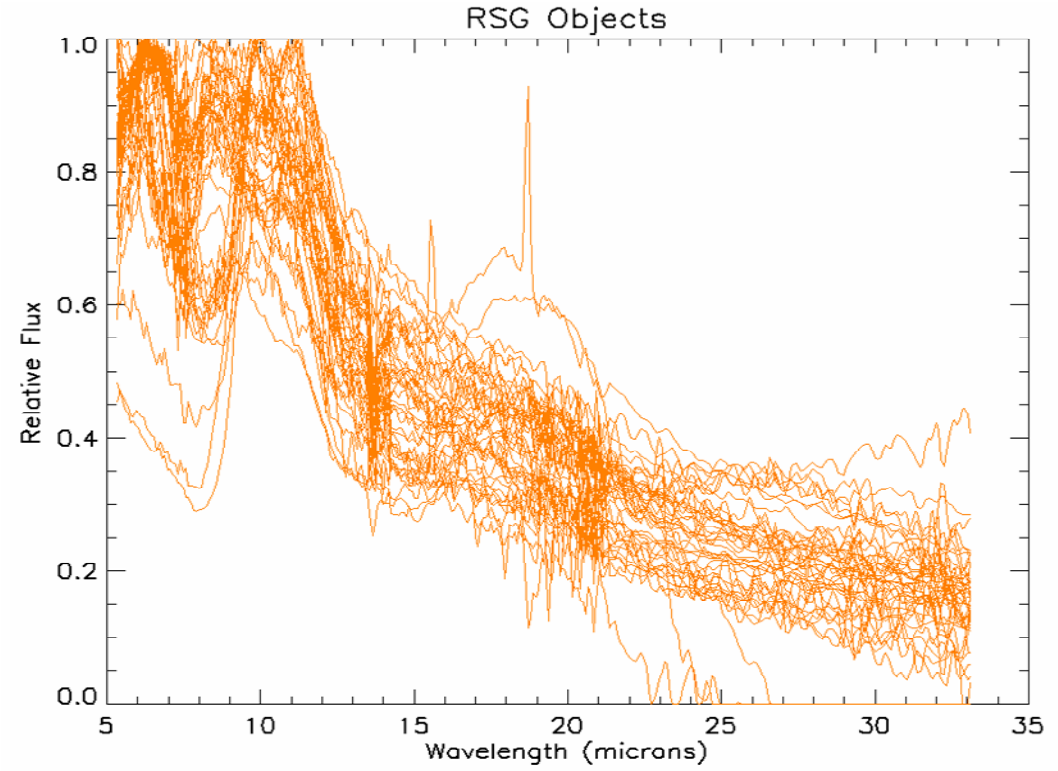

Figure 10. Spectra from the Sloan database classified as IRS type RSG.

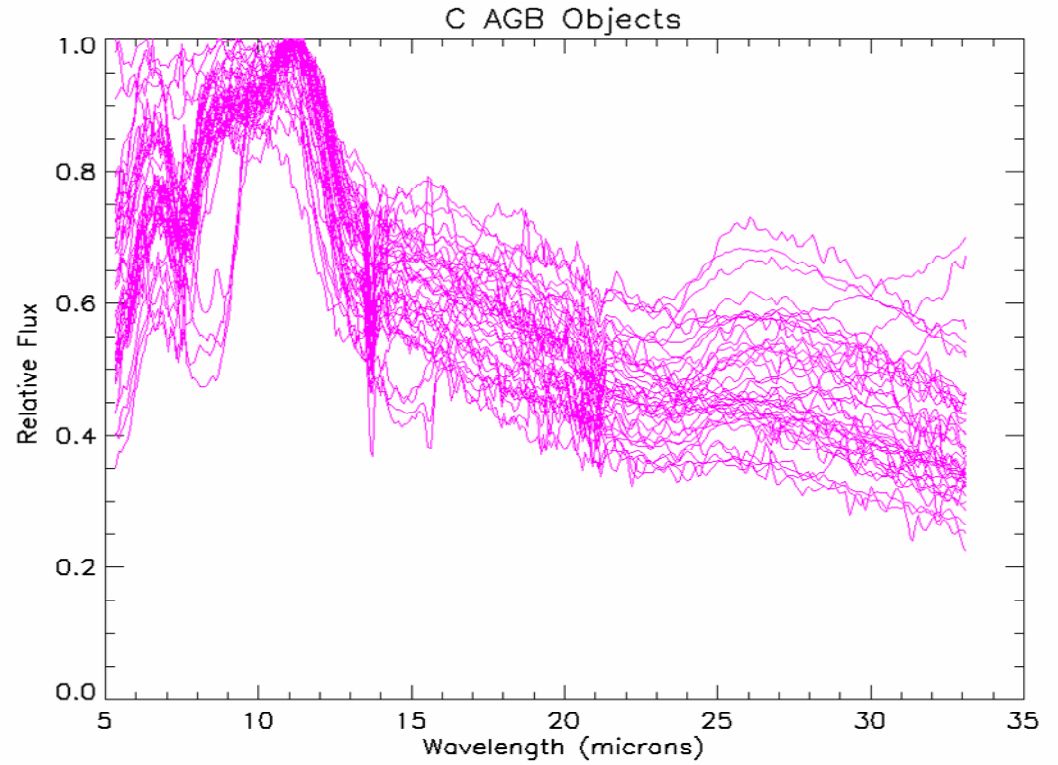

Figure 11. Spectra from the Sloan database classified as IRS type C AGB.

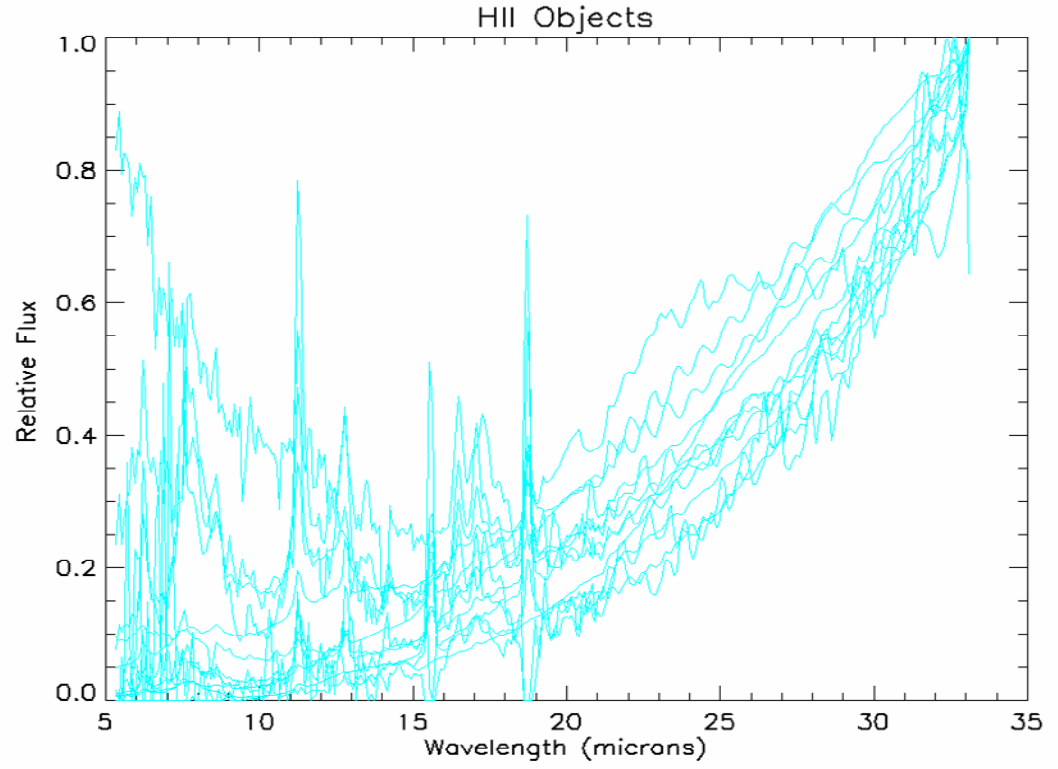

Figure 12. Spectra from the Sloan database classified as IRS type HII.

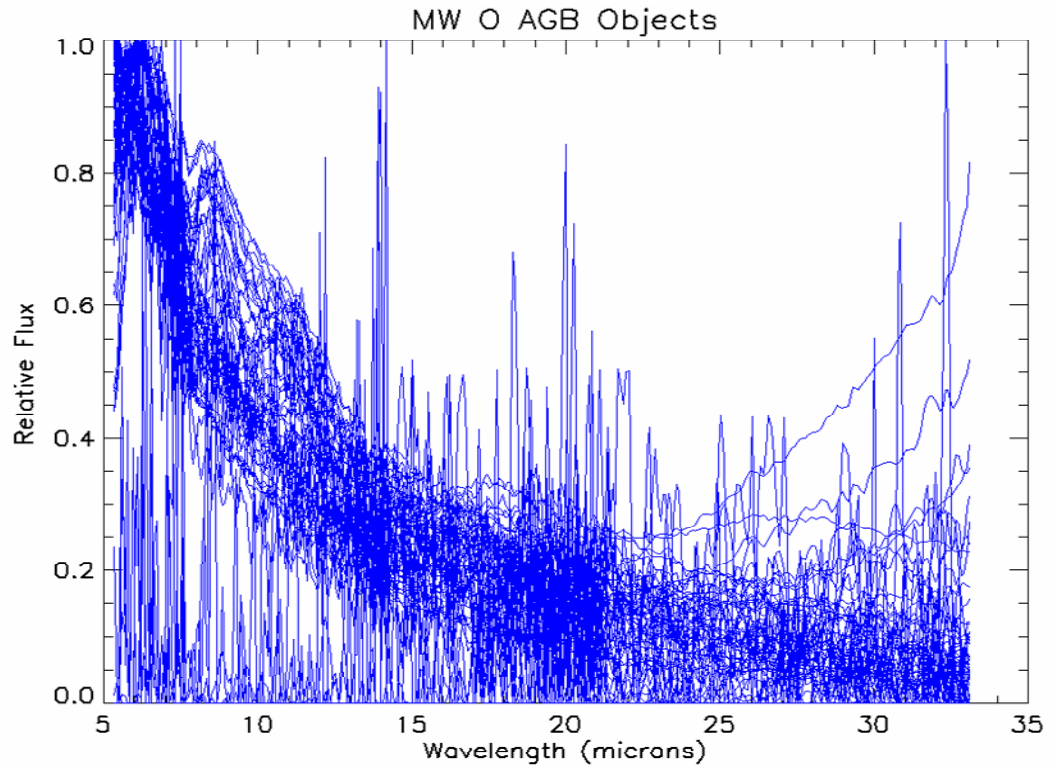

Figure 13. Spectra from the Sloan database classified as IRS type MW O AGB.

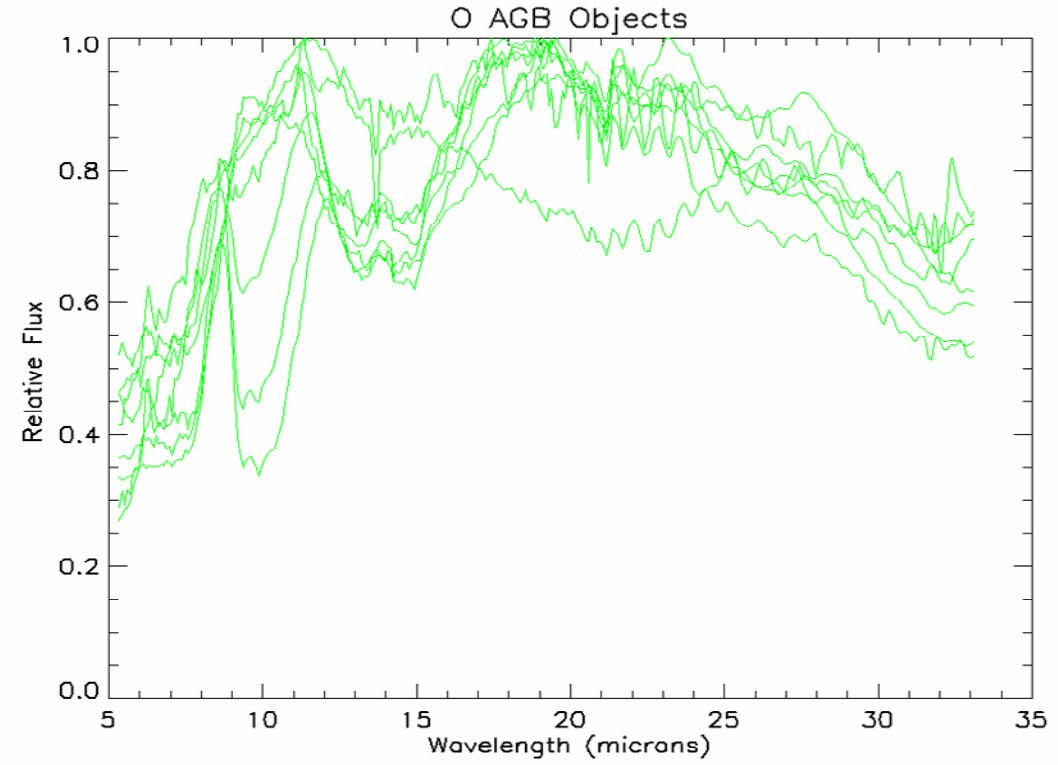

Figure 14. Spectra from the Sloan database classified as IRS type O AGB.

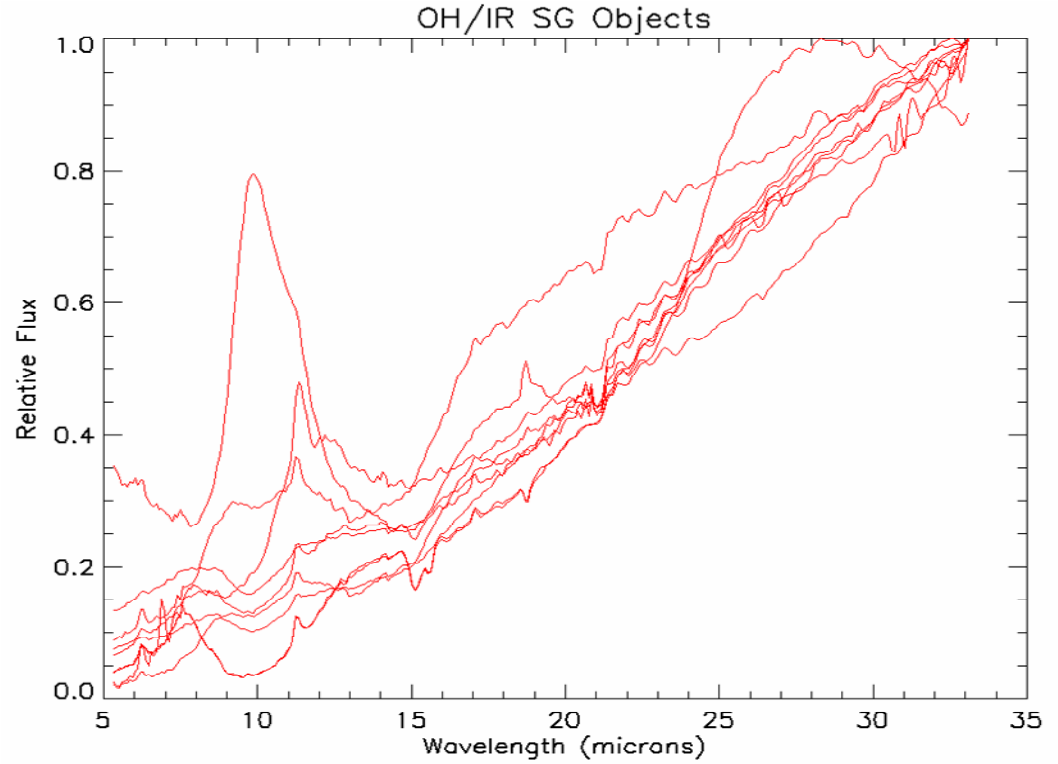

Figure 15. Spectra from the Sloan database classified as IRS type OH/IR SG.

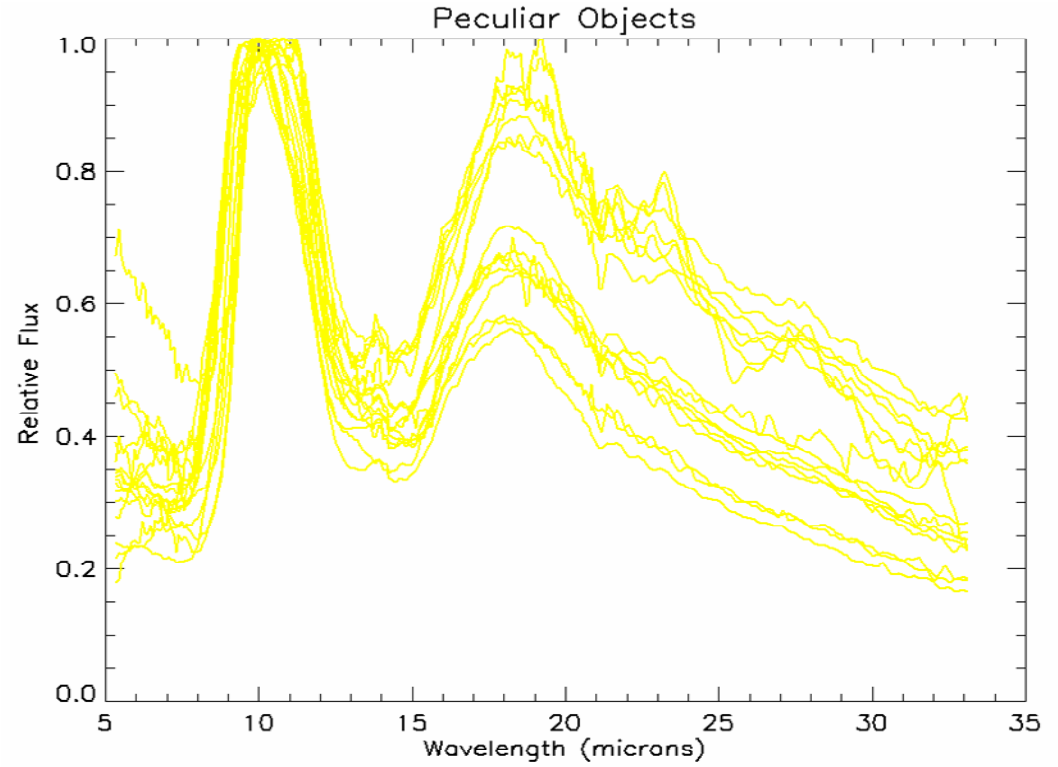

Figure 16. Spectra from the Sloan database classified as IRS type "peculiar."

Figures 9 through 16 illustrate that the spectral curves representing the IRS types are very similar between the Buchanan and Sloan databases. HII spectra remain very distinctive and appear to be classified correctly. Figure 15, however, indicates that some HII spectra may be intermixed with type MW O AGB objects. While there appear to be many MW O AGB spectra at low relative flux densities, Figure 13 indicates that spectral noise may have negatively impacted their classification.

There also appears to be a distinct separation between the C AGB and RSG objects within both databases. Spectra of these two groups often overlapped in the Buchanan database; the main difference being that RSG objects lie below C AGB objects with higher normalized flux within the 5µm to 10 µm range. This was not the case in the Sloan database. C AGB stars exhibited higher flux densities from 10µm to 33µm. Since the astronomical community has not yet reached a consensus as to the IRS types for all objects in the Sloan database, it is not possible

to calculate the classification accuracy. While we found that classification technique worked well to separate the spectra of the Sloan database, further work is needed to determine the classification accuracy.

FUTURE WORK

There are numerous areas where future work is needed for this experiment. First, methods to improve the overall classification accuracy for the Buchanan database may be examined. One possible method includes restricting space to a series of deterministically selected spectral regions to give more weight to specific features. Such a technique may provide a more accurate separation of MW O AGB, C AGB, and RSG objects.

Methods to reduce the dimensionality of the dataset may also be investigated. Principal Components Analysis (PCA) uses a principal components transform to generate a new set of components, or bands, that represent an alternative description of the data. The new components of a spectral vector are related to the original values by a linear operator. The important and essential information content of the spectral database are represented by a reduced number of transformed dimensions, thus allowing for classification to be conducted on a reduced number of bands (such as the first 3 bands, instead of all 325) while also increasing classification accuracy.

Alternatively, methods to reduce noise within the spectra may be examined. PCA often works well when there is minimal or uniform noise; for astronomical spectra, it may be helpful to remove the noise from the image before conducting PCA. The minimum noise fraction (MNF) technique allows for such noise reduction. MNF produces new components that exhibit decreasing spectral quality with increasing component number, and also maximize the signal-to-noise ratio. PCA, on the other hand, maximizes variance and does not always produce a decreasing spectral quality with increasing component number. MNF is implemented by removing the noise from the image, and then applying the principal components transform to the filtered image.

Lastly, future work is necessary to examine classification accuracy for the Sloan database. While these objects are presently under examination by the astronomical community, their corresponding IRS types have yet to be determined. While our technique provides an excellent foundation for classifying these objects, further examination is needed to determine how well our classification technique performs.

CONCLUSION

A classification scheme has been developed to classify astronomical spectra into subgroups characterized by the object's Spitzer IRS spectral type. This classification scheme operates by performing a hierarchical clustering routine on the Buchanan spectral database and separating the spectra into one of seven groups. Confirming a key result of Buchanan *et al.*, by examining a similarity curve we have also shown that seven is an optimal number of IRS spectral groups. In other words, we have shown that the level of similarity between successive clusters decreases significantly once seven clusters have been obtained.

Once we determined class membership for each object in the spectral database, we calculated the mean of the spectra in each class, and applied these spectral vectors as initial cluster means to a K-means clustering algorithm. The routine iterated eight times until no objects changed class membership. Analysis of the confusion matrix indicated an overall accuracy of 80%. Forty-one of fifty-one spectra were classified correctly, and all HII spectra were accurately grouped. However, some objects of MW O AGB, C AGB, and RSG types were classified incorrectly and mixed in with groups of other IRS types.

Using initial cluster means from the results of hierarchical clustering on the Buchanan database, we implemented the K-means clustering algorithm on the Sloan spectral database. While we found that the classification scheme works well to separate the objects of an unknown database into groups corresponding to their IRS types, further work is needed to fully examine the classification accuracy of this dataset. In conclusion, we have shown that classification of Spitzer IRS spectra is indeed possible, and that analysis and use of our clustering techniques will provide Spitzer astronomers and scientists with a valuable tool to classify large databases of unknown or known IRS spectra.

REFERENCES

Buchanan, C. L., and J. H. Kastner, et al... "A Spitzer Space Telescope Infrared Spectrograph Spectral Atlas of Luminous 8μm Sources in the Large Magellanic Cloud." The Astronomical Journal 132 (Nov. 2006): 1890-1909.

Duda, R. O., P. E. Hart, and D. G. Stork. "Unsupervised Learning and Clustering." Pattern Classification. 2nd ed. New York: John Wiley & Sons, Inc., 2001. 517-600.

Hojnacki, S. M., et al... "An X-ray Spectral Classification Algorithm with Application to Young Stellar Clusters." The Astronomical Journal 2007 (in press).

Hojnacki, S. M., and J. H. Kastner. "Automated Classification of X-ray Sources in Steller Clusters." Proceedings of SPIE 5493 (2004): 474-482.

Johnson, R. A., and D. W. Wichern. "Clustering, Distance Methods, and Ordination." Applied Multivariate Statistical Analysis. 5th ed. Upper Saddle River (NJ): Prentice-Hall, Inc., 2002. 668-747.

Life Cycle of a Star. University of Utah. 11 May 2006 <http://aspire.cosmic-ray.org/labs/star_life/starlife_main.html>.

Mu, B., J. H. Kastner, and C. L. Buchanan. "A Performance Comparison of Unsupervised Clustering Techniques for Classification of Spitzer Space Telescope Infrared Spectra." Center for Imaging Science, Rochester, NY.

Richards, J. A., and X. Jia. Remote Sensing Digital Image Analysis. 3rd ed. Berlin (Germany): Springer-Verlag Berlin Heidelberg, 1999. 133-148.

Sloan, G. "Spitzer Joint Magellanic Cloud Programs." Unpublished.

Wolf, C., K. Meisenheimer, and H. J. Roser. "Object Classification in Astronomical Multi-color Surveys." Astronomy and Astrophysics 365.3 (Jan. 2001): 660-80.