Exploratory Data Analysis in Multi-Dimensional Big Data

Brad Strylowski, Kirk Borne, Arun Vedachalam

College of Science COS Undergraduate Research URC Colloquium 2013

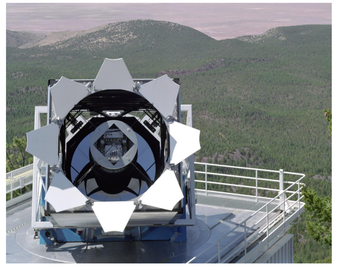
Abstract

In this project, we ran multi-dimensional cluster analysis on a set of galaxies in order to explore which combination of attributes (which sets of parameters) would best separate two different classes of galaxies: non-merging galaxies versus merging galaxies. Our dataset consisted of 2810 galaxies that had been verified by astronomers as mergers, and 3500 that had been similarly verified as non-mergers. Each galaxy is characterized by a set of 75 pre-selected attributes that were obtained without human observation (i.e., recorded through an automated image processing data pipeline). Cluster analysis was run on the dataset in combinations of 1, 2, 3, and 4 of these galaxy attributes. For each combination of attributes, the quality of the data clouds (merger and non-merger clusters) was measured by a cluster validation metric: the Davies-Bouldin Index (DBI: equal to the sum of the radii of the clouds divided by the separation of the clouds). A small value the index would indicate strongly compact clusters, thus revealing strong separation of mergers from non-mergers in that set of chosen parameters. After normalizing the full dataset's parameter values to a consistent scale, we used the standard deviation of a cloud's parameter values as its radius and the distance between the clouds' centroids as the cluster separation. We obtained our lowest values for the DBI (corresponding to the best separation between clouds) in combinations of one attribute. As we increased the dimensions (number of parameters) representing the data clouds, the lowest numeric value of DBI also increased, indicating that the merger and non-merger clouds were not separating as we had hoped. Nevertheless, certain attributes consistently appeared in the set of attributes with the best separation, indicating that merging galaxies may be slightly more likely than non-merging galaxies to have certain values of these parameters. These results will enable astronomers to automatically discover this special class of galaxies (mergers) within the massive astronomical data collections of the future, which will include tens of billions of galaxies.

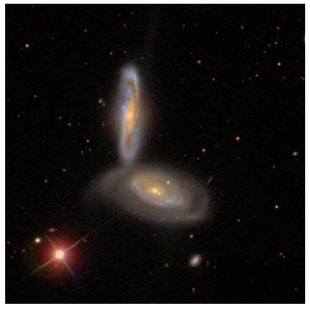
Introduction

In recent years, large astronomy projects like the Sloan Digital Sky Survey have offered scientists a wealth of data on galaxies consisting of both images and spectrums. However, identifying the relevant and interesting galaxies remains a challenge due to the massive quantity of data. In this project, we sought to distinguish merging galaxies from non-merging galaxies by data that can be automatically collected from the spectrum of galaxy, thus eliminating the need for a professional astronomer to examine the image. We used the Davies-Bouldin Index of the merger and non-merger clouds in several combinations of these automatically collected attributes to distinguish attributes that separated the clouds from parameters with large overlap.

Our dataset consisted of 6310 galaxies (the same set used in the "Data Mining the Galaxy Zoo Mergers" project), each of which had 75 attributes that were obtained from a computer analysis of the spectrum. These galaxies had all been classified by professional astronomers as either merging or non-merging. Before calculating the separation of the clouds, we normalized the data for each attribute by subtracting the mean value of the attribute and dividing by the standard deviation.



The telescope used for the Sloan Digital Sky Survey at Apache Point Observatory in New Mexico.



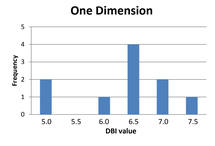
Two galaxies, NGC 3786 and NGC 3788, that are gravitationally interacting.

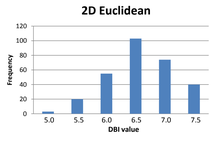
Methods

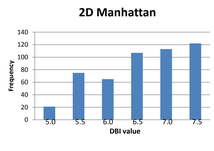
To distinguish mergers and non-mergers, we ran cluster analysis on our set of galaxies in four (out of 75 possible) dimensions. Cluster analysis seeks to separate objects of different classes by identifying different clusters of data points when the data is characterized by a few attributes. Our hope was that as we progressed into higher dimensions, the overlap apparent in each of the attributes would disappear as more attributes helped separate the mergers from the non-mergers.

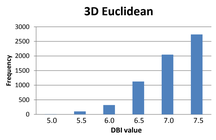
Although several measures of cloud separation exist, we chose to use the Davies-Bouldin Index (DBI) to measure the separation of the clouds. The DBI measures cloud separation by dividing the sum of the clouds' radii by the separation of the clouds. For our research, this meant the standard deviation of the merger and non-merger clouds divided by the distance between the clouds' centroids. We chose to use Euclidean distance for distance calculations in four dimensions, but we decided to use Manhattan distance for a separate trial in two dimensions. Note that attributes with traditionally larger values did not dominate distance calculations, as we normalized before cluster analysis. Compact clouds with large separation will produce low DBI values, while overlapping clouds with high deviation will produce high DBI values.

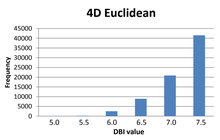
We used a set of three python programs to run cluster analysis in four dimensions. First, we tested the program in one dimension, and compared the results with manual DBI calculations made in Microsoft Excel. Then, using Euclidean distance, we expanded the number of dimensions (or combinations of attributes) to 2, 3, and 4, using Euclidean distance for distance calculations. For the fourth dimension, we created a threshold DBI value of 7.5, so as to only measure the best clouds (approximately the best 10% based on our previous results). We also calculated the DBIs for two dimensional clusters using Manhattan distance.



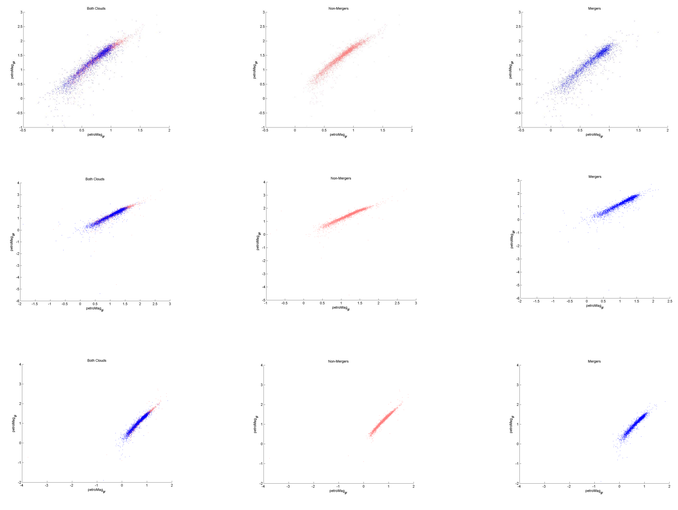








A histogram of the Davies-Bouldin indeces obtained in each of our trials. The bins are constructed to span the lower end of the spectrum of results.



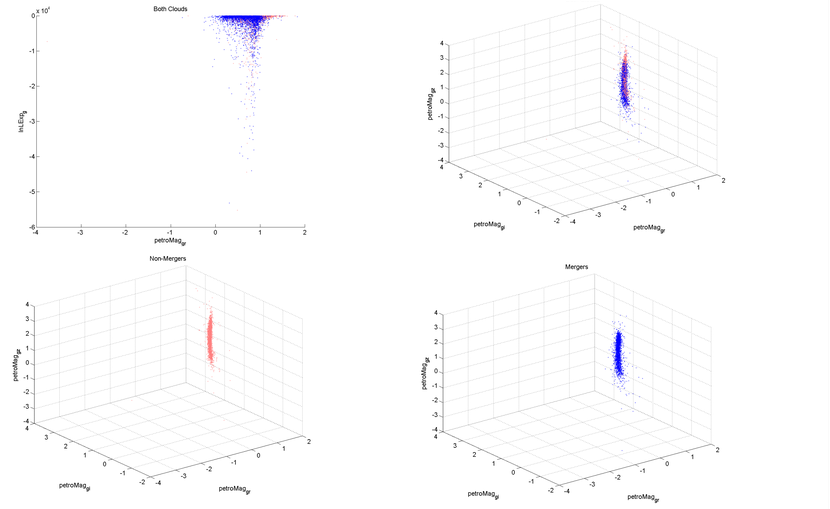
Results and Conclusions

As we progressed into higher dimensions, the merger and non-merger clouds did not separate as we had hoped. The best separation (using Euclidean distance) occurred in one dimension, with the petroMag_g_r attribute (Petrosian flux of red light subtracted from green light) producing a DBI of 4.43. Each time we added a dimension, our minimum DBI value increased, indicating that adding attributes caused more overlap instead of separation.

Although the data clouds did not separate, our results offered insight into distinguishing mergers and non-mergers. Several attributes contributed to lower DBI values across several dimensions, indicating that merging galaxies are more likely to have certain values of these attributes than non-merging galaxies. Most of these attributes that offer better separation are measures of blue or green light, indicating a difference in the amount of blue and green light coming from merging and non-merging galaxies.

After cluster analysis, we generated some two-dimensional scatterplots of the best combinations of attributes and discovered a possible oversight of our separation metric. For three of the four the best combination of attributes, the data clouds took the form of ellipses with high eccentricity. The Davies-Bouldin Index best measures the separation of clusters when the clusters are circular. Adjacent ellipses with a high eccentric spread will produce large DBI values even when the separation is perfect, because the radii of the clouds will be very large in comparison to the separation of the centers.

Our two-dimensional test run using Manhattan distance generated some interesting results. Our minimum DBI value actually decreased, indicating a better separation of clusters. Furthermore, the best combination of attributes in Manhattan distance (petroMag_g_r and lnLExp_g) was not the same as the best combination in Euclidean distance (petroMag_g_r and petroMag_g_z). Strangely, the fourth-best combination in the Euclidean trial offered a much better separation in the Manhattan trial.



Acknowledgements

Dr. Kirk Borne, Professor of Astrophysics and Computational Science, George Mason University
Arun Vedachalam, graduate student, George Mason University
The Galaxy Zoo Project
The Sloan Digital Sky Survey
David L. Davies
Donald W. Bouldin