SDSS MOC4 Asteroids' Color Classifications

Benjamin Montgomery
Department of Computer Science
University of Southern Maine
Portland, Maine 04104
Email: benjamin.montgomery@maine.edu

***Abstract*—The range of taxonomic categories of asteroids found to be present in the Sloan Digital Sky Survey (SDSS) has varied markedly within the span of a decade, bringing to question the particular methods used by machine learning algorithms to analyze and classify this data. We used the Agglomerative, HDBSCAN*, K-Means++, and K-Medoids clustering algorithms to attempt to independently verify the methods used by previous papers, finding that some previous classification efforts are not functions of color clusters.**

I. Introduction

The Sloan Digital Sky Survey is a multi-spectral survey of celestial objects, most notably of asteroids. Multiple efforts [1], [2], [3], [4] have been presented to classify these asteroids based on their observed spectra and colors; the expected number of classes existent in the data ranges from two classes [4] to sixteen [3]. These analyses fail to capture the accepted twenty-six [5] classes found using other asteroid spectral datasets. The variance appears largely due to the kind of classification methods used; original attempts were made using unsupervised clustering algorithms [2], [3], whereas later efforts were supervised learning algorithms that relied on human-specified templates [6] of what each class of asteroid could look like. This brings to question if similar results could be produced by an unsupervised learning algorithm using the modern color correction methods found in recent research, specifically that of Carvano [3].

II. Methods

We used data processing methods identical to that presented in Carvano's [3] work, namely, classification based on the SDSS' recorded values of the u', g', r', i', and z' filters, with the v band used for normalization. Similarly, we adopted the color correction steps made for finding reflectance [3] colors, and for normalizing each asteroid observation by its reflectance color gradients [3].

Unlike Carvano's probabilistic interpretation of this information for classification purposes, however, we made separate attempts to classify the data present using Python-compatible implementations of the Agglomerative, HDBSCAN*, K-Means++, and K-Medoids unsupervised learning algorithms. These algorithms were provided using the HDBSCAN [7], Pyclustering [8], and Sklearn [9] packages for the first three algorithms, and a custom vectorized implementation of the final algorithm. These algorithms were chosen due to the fact that they have been demonstrated to work well for other astronomical objects recorded through the SDSS' color filters, notably with galaxy classification [10].

Ideally, we could compare results to existing taxonomies, using tools such as the S-MASS II spectral survey, which includes asteroids classified using both the Tholen and Bus [11] taxonomies. Given that the intersection of asteroids in the SDSS MOC4 survey and those present in S-MASS II only has a cardinality of 90, it was necessary to employ alternative measures. The apparent best option was to compare each asteroid classification to that of Carvano's effort. Classifications were also checked across all observations to ensure that in every sighting of a given asteroid, classifications were consistent.

The SDSS is organized such that one row is equivalent to a single sighting of an arbitrary asteroid. All identifying names, along with observational data from the specific sighting, are included in these rows. Carvano's correction methods leave four columns valid for classification prediction; each represent normalized color gradients. These four calculated columns appear across the SDSS MOC4's 471,569 rows. Unsupervised clustering is performed on this dataset, and the results are combined in an outer join with Carvano's results, reducing the dataset down to 146,600 rows.

It is notable that no other correction effort was made, such as corrections for the zenith asteroids were observed at, or, more importantly, no efforts to fix the observed phase reddening [12] in this kind of astronomical data. Lack of these corrections causes significant spread of data in the u', g', r', i', and z' bands [12], casting doubt on all classification efforts. This apparent spread can be seen in the T-distributed Stochastic Neighbor Embedding (TSNE) and Principle Component Analysis (PCA) visualizations in Figure 1. In this figure, as with all dimensionality reduction figures in this paer, each color represents a different class.

III. Algorithm Results

Algorithm results were primarily analyzed by consistency, which is expressed for a given asteroid a out of n total, having m_a observations, and the number of class occurrences $c1_a, c2_a...cn_a$ for a given classification cn:

$$\sum_{a=0}^{n} \frac{max(c1_a, c2_a...cn_a)}{m_a}$$

By this metric, an asteroid classified with class 1 twice, class 2 three times, and class 5 one time has a consistency of $\frac{max(2,3,1)}{6} = 0.5$. This metric is useful when all classes are about the same size; else, it must be compared to blindly guessing the dominating class.

For comparisons with Carvano's work, bar graphs were created for each algorithm-generated cluster, representing the number of occurrences each of Carvano's classes appear within the cluster.

A. K-Means++

K-Means, and therefore K-Means++, is particularly susceptible to noise and outliers. Unfortunately, the SDSS is known to have a large quantity of outliers in its data, one author going as far to write "The lack of success [in previous classification attempts] is due to ... the relatively high uncertainties of the observations of the fainter objects, which dominate the sample." [13]. None of the aforementioned literature address the existence of possible distributions of data, making outlier recognition exceedingly difficult. Efforts to this end appear limited to aggressive data filtering [14] by the number of degrees from the galactic plane, something of a list-ditch attempt that we did not try to replicate.

With this limitation in mind, K-Means was used as a heuristic to inform guesses about the number of clusters. Two methods of predicting the number of clusters were used, and are detailed as follows.

1) Equation-Based Elbow Calculation: Given k_{min}-point (x_0, y_0) the k_{max}-point (x_1, y_1), the amount of clusters x_k, and the within-cluster error y_k, we may find the elbow point with the equation

$$Elbow_k = \frac{(y_0 - y_1)\, x_k + (x_1 - x_0)\, y_k + (x_0 y_1 - x_1 y_0)}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}}$$

The range $k \in \{i \mid i \in \mathbb{N}, (0 < i < 25)\}$ was used to evaluate this equation, reflecting the upper bound set by Binzal's classification efforts [5]. Using this approach, we concluded that, without removal of any possible outliers, there are nine distinct clusters.

2) Manual Elbow Calculation: As before, the range $k \in \{i \mid i \in \mathbb{N}, (0 < i < 25)\}$ was used, and for every k, K-means was run ten times to ensure that the initialization led to a global minimum. The resultant graph is detailed in Figure 4a. Using this approach, we concluded that, without removal of any possible outliers, there are seven to nine distinct clusters. This algorithm resulted in a consistency of 87.8%, meaningfully greater than the blind guess accuracy of 60.0%.

B. K-Mediods

Due to limitations in the efficiency of PyClustering's implementation of the K-Mediods algorithm, it was deemed necessary to implement one that effectively reflects the Partitioning Around Medoids (PAM) implementation. Given the freedom to write custom distance metrics, the median absolute deviation was chosen as the optimal comparison metric due to prior research indicating its superiority in this dataset [14]. This metric led an apparent elbow around $k = 9$ over the course of 10 runs of evaluating $k \in \{i \mid i \in \mathbb{N}, (0 < i < 25)\}$ as before, as shown in Figure 4b.

This algorithm resulted in a consistency of 77.3%, meaningfully greater than the blind guess accuracy of 25.7%.

C. HDBSCAN*

HDBSCAN* notably has mechanisms built in to guess what datapoints are outliers. In practice for this dataset, it resulted in classifying nearly 400,000 points as outliers, a decision somewhat understandable given the fact that this corresponds to a low-density region in the data. Using distance metrics that ignored this density, such as the Ward distance, resulted in classifications similar to that of G-Mode [14], complete with a main cluster of size 468,279. An example using the Ward distance is show in Figure 5a.

D. Agglomerative Clustering

The Ward distance was once again used to classify the data, with a dendrogram of a random sampling of 10% of the datapoints used to determine a distance cutoff. This algorithm resulted in a consistency of 99.8% versus a blind guess consistency of 99.3%, which is likely not a statistically significant difference.

In no algorithm was there a particular correlation between Carvano's classifications and those found with unsupervised learning. Instead, there appeared to be an approximately equal distribution of all of Carvano's classes in each unsupervised cluster, agnostic of algorithm. A sample of this tendency is shown in Figure 6.

IV. Conclusion

Given the corrections specified, Carvano's classifications do not appear to be functions of asteroid groupings, or at least not ones that PCA or TSNE dimensionality reduction could indicate. Nor does it appear to hold a meaningful relation to clusters found by unsupervised learning. These classifications appear to be at least in part affected by the considerable noise in the data, as well as a large amount of outliers.

It is unclear if unsupervised learning is a scientifically valid approach for composition classification of the SDSS. Nor is it clear if the frequently-occurring large cluster actually exists, which would support the Tholen classification's tendency to state most asteroids are C/S type, or if it simply reflects the failure of unsupervised learning in light of insufficiently corrected data.

If unsupervised learning is a valid approach, there are approximately 9 distinct compositions present. This a replication and improvement upon recent unsupervised learning efforts with the G-Mode algorithm [14], which encountered the same large cluster, but was unable to meaningfully cluster asteroids falling outside this cluster, organizing them into 58 classes.

Acknowledgment

The author would like to thank Professor Julie Ziffer of the University of Southern Maine's Department of Physics for her role as an acting consultant in the interpretation of astronomical data.

References

[1] C. R. Chapman, D. Morrison, and B. Zellner, "Surface properties of asteroids - A synthesis of polarimetry, radiometry, and spectrophotometry," *Icarus*, vol. 25, pp. 104–130, May 1975.

[2] V. Zappalà, P. Bendjoya, A. Cellino, P. Farinella, and C. Froeschlé, "Asteroid families: Search of a 12,487-asteroid sample using two different clustering techniques." *Icarus*, vol. 116, pp. 291–314, Aug. 1995.

[3] J. M. Carvano, P. H. Hasselmann, D. Lazzaro, and T. Moth-Diniz, "SDSS-based taxonomic classification and orbital distribution of main belt asteroids," *Astronomy and Physics*, 2009.

[4] Ž. Ivezić et al., "Solar system objects in the SDSS commissioning data," *The Astronomical Journal*, july 2001.

[5] S. J. Bus and R. P. Binzel, "Phase II of the Small Main-Belt Asteroid Spectroscopic Survey. A Feature-Based Taxonomy," *Icarus*, vol. 158, pp. 146–177, Jul. 2002.

[6] R. P. B. et al., "Searching for v-type and q-type main-belt asteroids based on sdss colors," *Lunar and Planetary Science*, 2007. [Online]. Available: https://www.lpi.usra.edu/meetings/lpsc2007/pdf/1851.pdf

[7] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[8] A. Novikov, "annoviko/pyclustering: pyclustering 0.8.1 release," May 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1254845

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

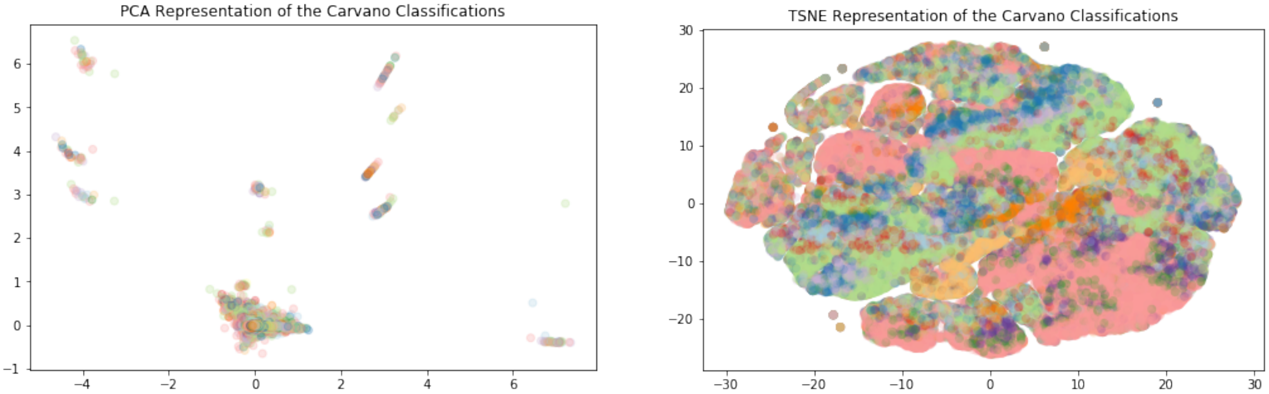
[10] J. Sánchez Almeida, J. A. L. Aguerri, C. Muñoz-Tuñón, and A. de Vicente, "Automatic Unsupervised Classification of All Sloan Digital Sky Survey Data Release 7 Galaxy Spectra," *American Journal of Physics*, vol. 714, pp. 487–504, May 2010.

[11] S. J. Bus, "Compositional structure in the asteroid belt: Results of a spectroscopic survey," 1999.

[12] J. A. S. et al., "Phase reddening on near-earth asteroids: Implications for mineralogical analysis, space weathering and taxonomic classification," *Icarus*, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0019103512001376?via\%3Dihub

[13] P. H. H. et al., "Characterizing spectral continuity in sdss ugriz asteroid photometry," *Astronomy & Astrophysics*, 2018. [Online]. Available: http://adsabs.harvard.edu/abs/2014acm..conf..206H

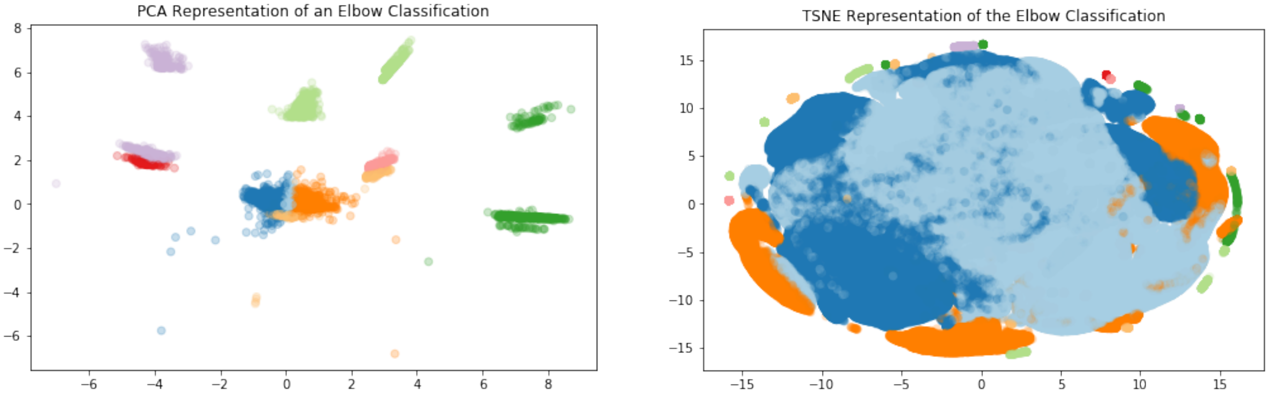
[14] ——, "Adapted g-mode clustering method applied to asteroid taxonomy," *PROC. OF THE 12th PYTHON IN SCIENCE CONFERENCE*, 2013.



(a) PCA Reduction

(b) TSNE Reduction

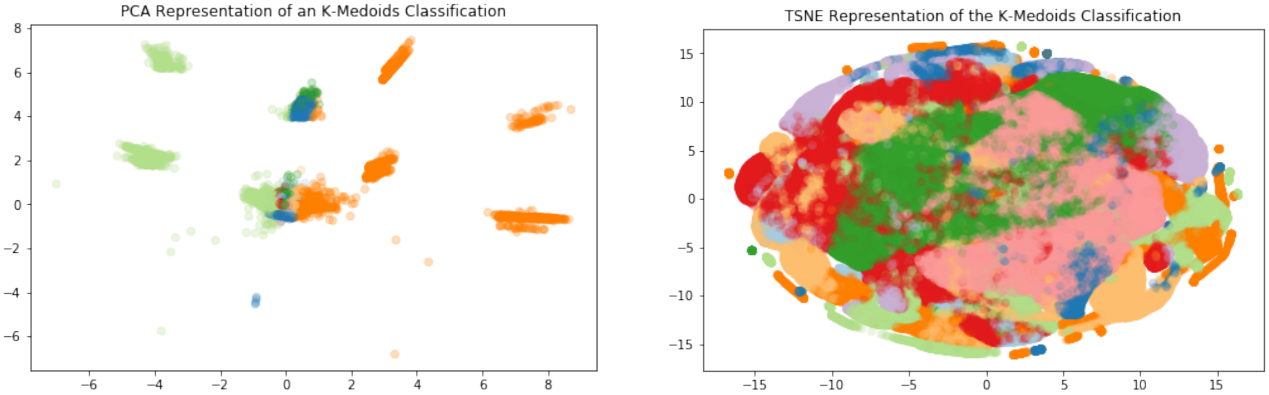
Fig. 1. Dimensionality Reduction of Carvano's UGRIZ-based Classifications



(a) PCA Reduction

(b) TSNE Reduction

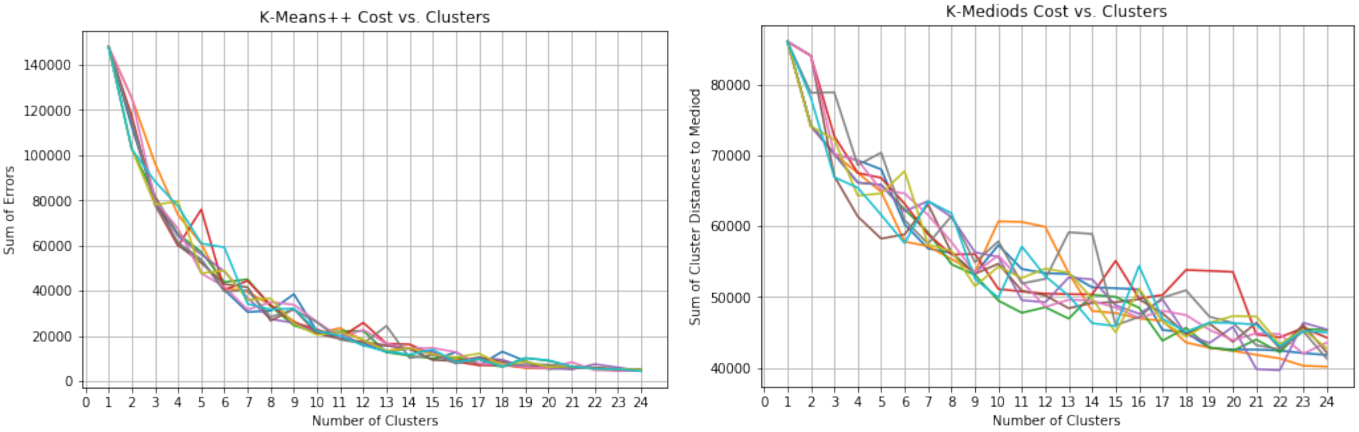
Fig. 2. Dimensionality Reduction of the Equation-Based Elbow Method with K-Means++



(a) PCA Reduction

(b) TSNE Reduction

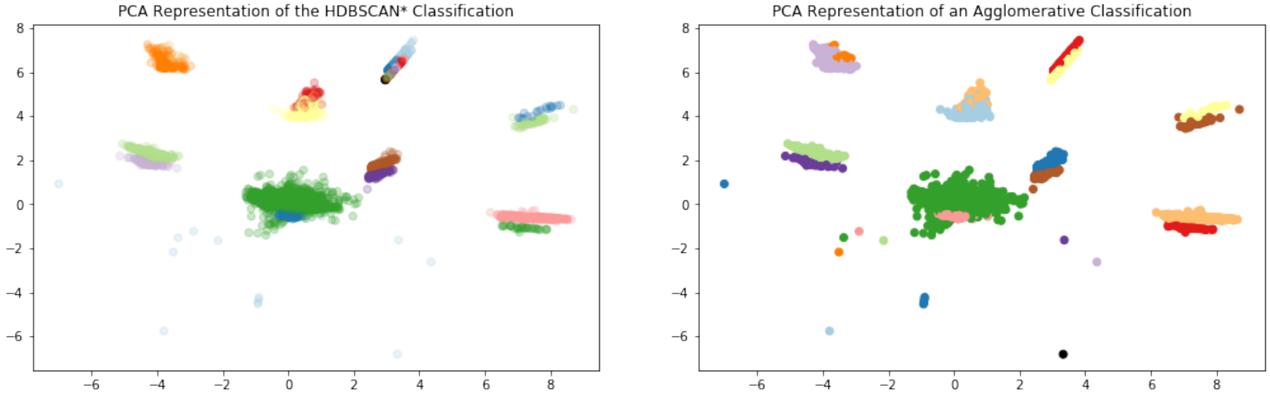
Fig. 3. Dimensionality Reduction of K-Medoids Clustering



(a) Graphical K-Means++ Elbow Method

(b) K-Medoids Cluster Selection

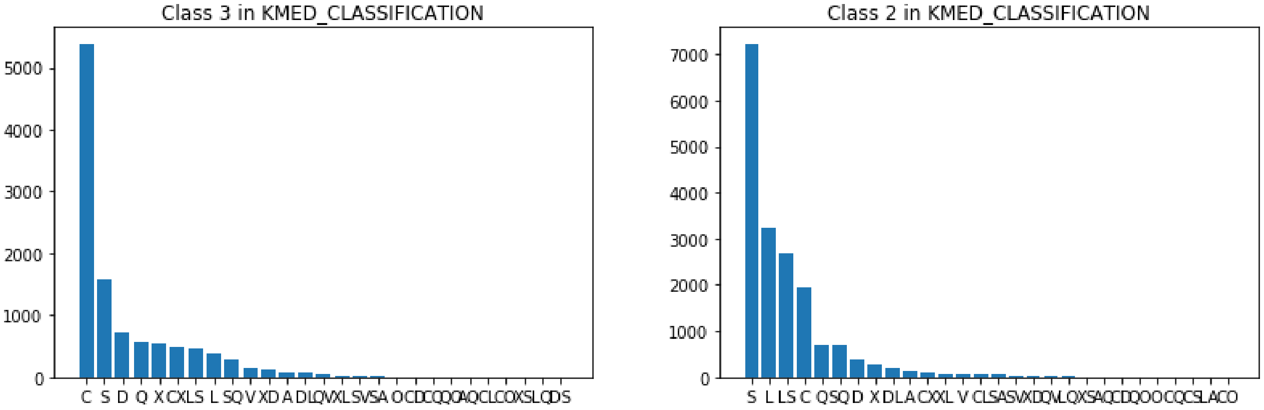
Fig. 4. Cost/Error Analysis for the K-Means++ Elbow Method, K-Medoids Cluster Selection.



(a) PCA Reduction of HDBSCAN*

(b) PCA Reduction of Agglomerative Clustering

Fig. 5. Dimensionality Reduction of the Equation-Based Elbow Method with K-Means++



(b)

Fig. 6. Sample bar Graphs Against Carvano with K-Medoids