Proceedings of the International Meeting on “High-Dimensional Data-Driven Science” (HD3-2017)
September 10–13, 2017, Kyoto, Japan

Taxonomy matching between asteroids and meteorites: supervised clustering approach

Yuki Saito∗, Peng Hong†, Takafumi Niihara‡, Hideaki Miyamoto†, Kenji Fukumizu§∗

∗Department of Statistical Science, School of Multidisciplinary Sciences, Graduate University for Advanced Studies, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan; †Department of Systems Innovation, School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan; ‡University Museum, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan; §The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

E-mail: saito.yuki@ism.ac.jp; hong@seed.um.u-tokyo.ac.jp; niihara@seed.um.u-tokyo.ac.jp; hm@sys.t.u-tokyo.ac.jp; fukumizu@ism.ac.jp

Abstract. We propose a data-driven method for investigating taxonomy relations between asteroids and meteorite. The method clusters data of reflectance spectra of asteroids with the guidance of relatively reliable taxonomy of meteorites. By comparing the clustering accuracy of asteroid between with and without guidance, we observe that the guidance of meteorites taxonomy improve the accuracy. This serves as an evidence that there is a common taxonomical structure between meteorites and asteroids, confirming a long-standing prediction.

1. Introduction

Linking meteorites to asteroidal bodies is an important subject toward better understanding of the Solar System. This is, however, not easy for various reasons such as space weathering. Some strong links between specific types are known so far. A link was confirmed between V-type asteroids and Howardite-Eucrite-Diogenite (HED) meteorites by the Dawn mission [2]. The link between S-type asteroids and ordinary chondrite meteorites was verified by the analysis of the sample taken from Itokawa by the Hayabusa mission [5]. A less certain link is known between C-type asteroids and carbonaceous chondrites. Other more uncertain links have been discussed also [3]. For investigating these links, there are some works that take data-driven approaches. Britt et al [1], Tholen et al [9], and Watabiki et al [10], for example, provided a map to overview the similarity in reflectance spectra between meteorites and asteroids by using principal component analysis.

This work proposes a data-driven method for linking the taxonomies between meteorites and asteroids. The proposed method classifies asteroid data with guidance of the known taxa of meteorites [11], and matches the taxa between the two domains of asteroids and meteorites. For this clustering and matching procedure, reflectance spectral data of asteroids are clustered so that the similarity between the cluster structures of the asteroids and meteorites is maximized. To represent the cluster structures, a nonlinear method, the so-called kernel method [7], is used to quantify the mutual relations among clusters. For the teaching data to guide the clustering of
asteroid data, the reflectance spectral data and element composite data of meteorites are used. We call this analysis procedure supervised clustering.

As a first step, we confine our analysis to the three types in each of the asteroids and meteorites: C-, S-, and V-type for asteroids, and the carbonaceous chondrites, ordinary chondrites, and HED for meteorites. The reflectance spectral data are used for asteroids, while for meteorites each of the reflectance spectra and element composites is served for teaching data.

2. Method for supervised clustering

2.1. Data sets

The data set of reflectance spectra for asteroid contains 365 data from 0.45 to 2.45 µm in wavelength divided by 401 bins of equal size. The dataset includes also the labels for the taxonomical types including C-, S-, V-, which are given by domain experts. These labels are not used for clustering but only for evaluation. As pre-processing, DeMeo’s method [4] is applied. The numbers of C-, S-, and V-type data are 19, 95, and 8, respectively; 122 in total.

For meteorites, the dataset of reflectance spectra contains 731 data, in which 221 carbonaceous chondrites, 245 ordinary chondrites, 108 HED meteorites, 574 in total, are used for the analysis. The preprocessing is the same as the asteroid data. The element composition dataset of meteorites consists of 481 data, consisting of 30 carbonaceous, 388 ordinary and 63 HED meteorites. The dataset includes 12 elements: Fe, Si, Al, Ca, Mg, Na, P, K, Mn, Ni, Cr, and Na₂O.

2.2. Supervised clustering

Our method for supervised clustering consists of two steps: [1st step] as initialization, the spectral clustering [6], which is a standard non-hierarchical clustering method, is applied to asteroid data, [2nd step] the meteorite data with taxonomic information is used to guide the clustering procedure for the asteroid data. We also refer to the standard clustering in the first step as unsupervised clustering in contrast with supervised clustering.

We review the spectral clustering in the sequel, since it serves as a basics of our supervised clustering. The spectral clustering operates on a similarity matrix, which contains as an element a similarity between two data points x and y. In this paper, the similarity is evaluated by Gaussian kernel function \( \exp(-\|x - y\|^2/2\sigma^2) \) with bandwidth parameter \( \sigma \). In the case of asteroid data, \( N_A \times N_A \) data-similarity-matrix \( K_A \) is used, which contains similarity between each pair of spectral data, where \( N_A \) is the size of the data set. The algorithm computes eigenvectors of the normalized graph Laplacian, defined by

\[
L_A := T_A^{-1/2}(T_A - K_A)T_A^{-1/2},
\]

where \( T_A \) is the diagonal matrix \( T_A,ii = \sum_j K_A,ij \), representing the degree of the nodes in the graph represented by \( K_A \). We used the algorithm proposed by Ng et al [6], which further applies K-means clustering after projecting the data onto the eigenspaces corresponding to the least eigenvalues.

The proposed method of supervised clustering is also based on similarity matrices. As an overview, the method first constructs a data-similarity-matrix for each of the asteroid and meteorite data, and then computes respective cluster-similarity-matrices for each domain to represent the cluster structures. A similarity measure between these two cluster-similarity-matrices is introduced for evaluating the similarity of taxonomies, serving as an objective function for clustering asteroid data and matching the clusters of the two domains.

For more details, we explain for the asteroid data. Given data-similarity-matrix \( K_A \) computed in the same way as spectral clustering, the cluster-similarity matrix is calculated by

\[
S_A = D_A W_A^T K_A W_A D_A,
\]
where $W_A$ is $N_A \times C$ cluster assignment matrix with common cluster size $C$ for the asteroid and meteorite domains, and $D_A$ is the normalizer. More specifically, $W_{A,ia} = 1$ if and only if $i$-th data belongs to $a$-th cluster with the other components 0, and $C \times C$ matrix $D_A$ is given by $D_{A,ab} = \delta_{ab}/N_A^a$ ($a = 1, \ldots, C$), where $N_A^a$ is the number of data in the $a$-th cluster $C_A^a$ of the asteroid dataset. The $(a, b)$-component of the cluster-similarity-matrix is then equal to

$$S_{A,ab} = \frac{1}{N_A^a N_A^b} \sum_{i \in C_A^a, j \in C_A^b} K_{A,ij},$$

which represents the average total similarities between the data in the two clusters. The cluster-similarity-matrix $S_M$ for the meteorite dataset is obtained in a similar way. Note that different data-formats, dimensionalities, and data sizes can be handled by the proposed similarity representations; the meteorite and asteroid datasets can have different data formats such as reflectance spectra and element composites.

The proposed objective function for our supervised clustering is to maximize

$$\frac{\text{Tr}(S_A H S_M H)}{\|H D_A W_A^T W_A D_A H\|_F} = \frac{\text{Tr}(H W_A^T D_A K_A D_A W_A H S_M)}{\|H D_A W_A^T W_A D_A H\|_F},$$

where $H$ is the centering matrix defined by $H_{ij} = \delta_{ij} - 1/C$, and $\|M\|_F$ denotes the standard Frobenius (Euclidean) norm. To optimize $W_A$, a greedy search is used in a similar way to CLUHSIC [8]. Before the search, $W_A$ is initialized by the spectral clustering, $W_M$ is obtained by the labels given by experts, and the cluster order is also optimized so that Eq.(4) is maximized.

3. Clustering results

For evaluation, the cluster purity and cluster matching accuracy are computed based on the type information labeled by experts for the asteroid data. The cluster purity of a clustering is the portion of data that accord with a dominant type in a cluster. The cluster matching accuracy is the portion of the correct cluster assignments to the corresponding meteorite type. This is similar to classification accuracy. In the following results, the best bandwidth parameters $\sigma_A$ (asteroid) and $\sigma_M$ (meteorite) are chosen, but from the experimental results not shown here, similar results are obtained by a wide range of the parameters.

Figure 1 shows supervised clustering of the asteroid data set with guidance of meteorite reflectance spectra data ($\sigma_A = 3.0$, $\sigma_M = 1.0$) (left) together with the teaching data of meteorites (right) for reference. They are shown after projecting onto the three dimensional space given by the principal component analysis. The cluster purity is 99.1% at best ($\sigma_A = 2.5$, $\sigma_M = 1.0$); this means significant improvement from the the cluster purity of the unsupervised clustering (spectral clustering), which is 92.8% at best ($\sigma_A = 1.5$).

As for cluster matching accuracy, the supervised clustering gives 99.0%. In Figure 1, the teaching cluster labels (right) are shown by A (carbonaceous chondrite), O (ordinary chondrite), and H (HED), while in the left figure C, S, and V show the labels given by domain experts (not used for the supervised clustering). The colors indicate the matched clusters; e.g. the red cluster A carbonaceous chondrite in the right figure is matched to red cluster in the left figure.

Next, we show supervised clustering results with guidance of element composites dataset for meteorites. Figure 2 is made in a similar manner to Figure 1. The cluster purity is 98.8% ($\sigma_A = 2.5$, $\sigma_M = 1.5$), which is significantly improved from the unsupervised clustering. The cluster matching accuracy is 98.9% ($\sigma_A = 2.5$, $\sigma_M = 2.0$) at best. It is important to note that the proposed supervised clustering works for different data types: reflectance spectra and element composites.

Those analyses show that taxonomic guidance of meteorite data in fact improves making asteroidal clusters, implying that a common cluster structure exists between the meteorite and asteroid domains.
4. Conclusion
We have studied a common taxonomical structure in asteroids and meteorites using a data-driven approach. We have proposed supervised clustering to approach this problem: it is investigated whether the taxonomy information of meteorite improves clustering of asteroid data. The numerical studies have demonstrated that the taxonomical information of meteorites does improve the accuracy of clustering of asteroids. This serves as an evidence that there is a common taxonomical structure between these two domains.

![Clustering of asteroid with guidance of meteorite reflectance spectra. Three dimensional plots by PCA for the clustering results of asteroid data (left) and the teaching data of meteorite (right).](image1)

**Figure 1.** Clustering of asteroid with guidance of meteorite reflectance spectra. Three dimensional plots by PCA for the clustering results of asteroid data (left) and the teaching data of meteorite (right).

![Clustering of asteroid with guidance of meteorite element compositions. Three dimensional plots by PCA for the clustering results of asteroid data (left) and the teaching data of meteorite (right).](image2)

**Figure 2.** Clustering of asteroid with guidance of meteorite element compositions. Three dimensional plots by PCA for the clustering results of asteroid data (left) and the teaching data of meteorite (right).

**Acknowledgments**
The present study was supported by KAKENHI Innovative Areas 25120012.

**References**
[1] Britt D T, Tholen J D, Bell J F, and Pieters M C 1992. *Icarus* **99** 153.
[2] Buratti J B, Dalba A P, Hicks D M et al. 2013. *J. Geophys. Res. Planets* **118** 1991.
[3] Cloutis A E, Binzel P R, and Gaffey J M 2014 *Elements* **10** 25.
[4] DeMeo E F, Binzel P R, Slivan M S, and Bus J S 2009 *Icarus* **202** 160.
[5] Nakamura T, Noguchi T, Tanaka M, Zolensky E M et al. 2011 *Science* **333** 1113.
[6] Ng A, Jordan M, and Weiss Y 2002. *Advances in NIPS* **14** 849.
[7] Schölkopf B and Smola A J 2001. *Learning with Kernels*. MIT Press.
[8] Song L, Smola A, Gretton A, and Borgwardt M K 2007 *Proc. ICML2007* 815.
[9] Tholen J D 1984 *Doctoral thesis, Univ. Arizona*.
[10] Watabiki Y, Nihara T, Hong P, Miyamoto H, Saito Y, and Fukumizu K 2016 *Japan Geoscience Union Meeting* MTT27-03.
[11] Weisberg K M, McCoy J T and Krot N A 2006 *Meteorites and the Early Solar System II*. 